Dataverse Community Survey 2022 – Report

Philipp Conzett

World map indicating the countries of Dataverse installations participating in the survey. The different colours indicate the numbers of Dataverse installations represented in the survey. The map was created with mapchart.net and is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

Table of Contents

Disclaimer

The author is a member of the Dataverse community, chair of the Steering Committee of the Global Dataverse Community Consortium (GDCC), and the repository manager of DataverseNO.

Acknowledgments

The author would like to thank the following members of the Dataverse community for providing valuable input to and feedback on early versions of the questionnaire of the Dataverse Community Survey 2022:s

Name	Affiliation	ORCID
Oliver Bertuch	Forschungszentrum Jülich	0000-0002-2702-3419
Jonathan Crabtree	University of North Carolina at Chapel Hill	0000-0002-0139-7025
Philip Durbin	Harvard University	0000-0002-9528-9470
Wim Hugo	Data Archiving and Networked Services (DANS)	0000-0002-0255-5101
Sherry Lake	University of Virginia	0000-0002-5660-2970
Courtney Mumma	University of Texas at Austin	0000-0003-1394-5006
James Myers	Global Dataverse Community Consortium (GDCC)	0000-0001-8462-650X
Kaitlin Newson	University of Toronto	0000-0001-8739-5823
Tania Schlatter	Harvard University	0000-0001-6857-659X
Don Sizemore	University of North Carolina at Chapel Hill	0000-0002-3761-2014

Summary

This report presents some of the results from the Dataverse Community Survey 2022.

Participants

With a total of 34 submitted responses from a variety of Dataverse installations, the survey participants reflect a large part of existing Dataverse installations. Two responses represent Dataverse installations that are in the planning phase.

Current status

Most of the participating Dataverse installations have an institutional focus, but also installations with a domain focus and global focus are represented in the survey. There are several installations that have a multi-institutional organization, including national, regional, and thematic consortia. Among the domain-specific installations, Social Sciences is the most common discipline being served.

Most of the Dataverse installations use a recent version of the main distribution (74%) of the Dataverse software that is manually (80%) deployed on local server(s) (67%). Installations on local server(s) using classic deployment methods tend to be deployed manually, whereas installations located in a cloud infrastructure tend to be deployed with containers using one or more deployment tools. The installations use local file storage (62%) and/or cloud-based file storage (47%). All installations use or are planning to use one or more configurable/optional features or tools within the following areas of usage:

- Getting Data In. Most popular: API tools; DVUploader; GitHub
- API Client Libraries. Most popular: Curl, wget, httpie or similar; pyDataverse; Dataverse R Package
- Embedding Data on Websites. Most popular: Widgets
- Analysis and Computation. Most popular: File Previewers; Data Explorer; Binder
- Discoverability. Most popular: Metadata harvesting of sources outside the installation through OAI-PMH; Collection and dataset linking; Google Dataset Search (.httpaccess set up for crawling)
- Curation and Preservation. Most popular: Submit for review configuration; BagIt export; Data Curation Tool
- File Management. Most popular: Tabular data file ingest; Embargo; Direct file upload
- Metadata Management. Most popular: Additional or customized metadata schemas; Multi-license configuration; External controlled vocabularies

- Other. Most popular: Branding: Header/footer/stylesheet; Branding: Customized home page; Dataset Guestbooks

A large majority of the Dataverse installations participating in the survey are already a member or are planning to become or considering becoming a member of the GDCC. Of the 34 Dataverse installations participating in the survey, 8 have contributed with software code, 10 have contributed to other community efforts, and 7 have contributed with both software code and other efforts.

Roadmaps and priorities

More than two thirds of the Dataverse installations participating in the survey either have, are planning to have or considering having a roadmap or strategic plan. To get an overview of the priorities of the installations, the survey included multiple questions about different areas of support provided by 1) the Dataverse software and 2) the Dataverse community / the Global Dataverse Community Consortium (GDCC).

1) Support provided by the Dataverse software

Most installations consider it as important or very important that the Dataverse software support the alignment or compliance with a number of requirements or recommendations, the most important ones coming from the following areas: FAIR principles, generic metadata harvesters, CoreTrustSeal, national stakeholders, and institutional stakeholders.

The respondents were also asked to rate the importance and indicate their preference for different architecture choices for 54 repository features within the following 14 topics:

- Core functionality
- RDM planning
- Deposit and Curation
- Large Data Support
- Metadata Support
- File Management and Reuse Support
- Data Privacy Support
- Documentation and Analysis Support
- Preservation Support
- Discovery and Reuse Support
- Accessibility Support
- Internationalization Support
- DevOps Support
- Scalability Support

Focusing on the top-five popular topics, the rating of importance can be summarized as follows: The repository features supporting core functionality are the ones that are considered as important or very important by the highest percentage of respondents (93%), followed by the features supporting deposit and curation (88% of the respondents), the features within the topic Metadata Support (78%), the feature supporting accessibility (77%), and the features supporting scalability (70%). Support for large data is considered less important as the file size and the number of files per dataset increase.

Of the 14 repository feature topics, core functionality is the one for which out of the box is the most preferred architecture choice (67% of the respondents), followed by Accessibility Support (55%), Internationalization Support (45%), Deposit and Curation Support (44%), Scalability Support (40%), DevOps support (37%), and Large Data Support (33%). For all other topics of repository features, extension is the most preferred architecture choice.

The Dataverse software guides and the guidance on best practice for data archiving and publishing work fine for most installations.

2) Support provided by the Dataverse community / GDCC

Almost all Dataverse installations consider it as important or very important that the Dataverse community (including GDCC) support the sustainable management and further development of the Dataverse software and the larger ecosystem of integrated services and tools.

More than half of the installations also rated the following areas of human/organizational support as important or very important:

- Support for install/administration (25 installations)
- Software installation/migration and upgrade (23 installations)
- Speakers who can represent the Dataverse community at conferences (19 installations)
- Repository certification (18 installations)
- PID providing (18 installations)

Virtually all installations consider the commitment by Harvard University and the commitment channelled through the GDCC as important or very important to ensure sustainable development and maintenance of the Dataverse software and the ecosystem of associated services and tools.

Both the project lead by Harvard and community governance support provided by GDCC are considered as important or very important by a large majority of the Dataverse installations participating in the survey. An even larger majority of the respondents reckon

that a transition of the GDCC to a non-profit organization will make GDCC better suited to provide governance support to the community. Coordination of community efforts and sustaining the Dataverse software and associated services and tools are seen as the two most important areas of governance support from GDCC.

Background

The Dataverse Project is an open-source web application to share, preserve, cite, explore, and analyse research data (King, 2007; *The Dataverse Project*, n.d.). The Dataverse software supports the archiving and publishing of research data according to the FAIR principles (Conzett, 2020, pp. 89–99; Crosas, 2020). The Dataverse Project is being developed at the Institute for Quantitative Social Science (IQSS) at Harvard University, along with a growing community of many collaborators and contributors worldwide.

Being a community-driven open-source software (OSS), The Dataverse Project faces many of the opportunities and challenges that are common in the OSS world. The importance of community for the sustainability of OSS projects is appropriately summarized in the report from a recent project on the topic:

> *[The] community and its diversity is a critical factor in the long-term sustainability of OSS, ensuring the software's ability to upgrade, adapt and grow to meet new needs and evolve with advances in technology (Arp & Forbes, 2022, p. 6).*

To support the coordination of community efforts, The Global Dataverse Community Consortium (GDCC) was established in 2018 with the vision of providing international organization to existing community efforts and providing collaborative venues for institutions to leverage economies of scale in support of Dataverse repositories around the world (Conzett et al., 2022; *GDCC*, n.d.).

Survey Objectives

The Dataverse software is continuously gaining popularity across the globe, and the Dataverse user and developer community has been growing correspondingly in the past years. This growth also brings with it some challenges when it comes to the question of how to organize community efforts for the best of the community. Addressing such challenges is one of the core tasks of the Global Dataverse Community Consortium (GDCC). To meet these challenges of growth, the Steering Committee of the GDCC have recognized the need for strengthening the sustainable development of the Dataverse software and its ecosystem of associated tools and services. To this end, GDCC are investigating how to further empower the GDCC to contribute to the Dataverse community with the support and services that are needed.

As a first step in this work, GDCC called earlier in 2022 on the Dataverse community to participate in a community survey to map out the current status as well as the roadmaps and priorities of Dataverse installations around the world.

The main goal of the Dataverse Community Survey 2022 is to help GDCC and the Dataverse Project decide on what actions to take to improve the Dataverse software and the larger ecosystem of integrated tools and services as well as better support community members.

The results from the survey may also be of interest to other communities working on software and services for managing research data.

Participants

The main target group for participating in this survey were the people/teams responsible for operating Dataverse installations around the world. A secondary target group were people/teams at organizations that are planning to deploy or considering deploying a Dataverse installation. There were 34 existing and planned Dataverse installations participating in the survey. For a list of the participating installations, see the presentation of the results from Question 2 in the *Results* section below. For more information about how Dataverse installations were invited to participate in the survey, see the *Methodology* section below.

Methodology

Survey design

The survey was designed to map out the current status as well as the roadmaps and priorities of Dataverse installations around the world.

The survey was organized in the following six main sections:

- An introduction containing the following:
    - Brief information about the survey
    - A request to respondents to submit only one coordinated answer for each Dataverse installation
    - Guidance on how to download the survey questions for those who wanted to have a look at the questions before answering the survey
    - A time estimate for completing the survey (1-2 hours)
    - Guidance on how to save the submitted answers before submitting the whole survey
- PART 1: DATA PRIVACY AND CONTACT INFORMATION containing the following:
    - Explanation of the purpose of the survey
    - Explanation of why the respondent is asked for consent
    - Information about the responsible institution (data controller)

- o Explanation of why the respondent is asked to participate
- o Explanation of what participating in the survey involves for the respondent
- o Informing the respondent that participation is voluntary
- o Information about how personal data will be stored and used, and further processed
- o Information about what rights the respondent has
- o Information about how to get more information about the survey
- o Granular consent to 1) participation; 2) processing of personal data; 3) open publication of non-personal data; 4) being contacted for follow-up questions
- PART 2: RELATION TO DATAVERSE PROJECT containing a question about the relation/interest of the Dataverse installation in the Dataverse community
- PART 3: INSTALLATION INFORMATION containing 30 questions about the Dataverse installations participating in the survey. The first 6 questions ask for general information about the installations, whereas the remaining 24 questions cover the following topics:
  - o Installation Type
  - o Organization
  - o Deployment
  - o Configurable/Optional Features used by the installation
  - o Community Engagement
- PART 4: ROADMAP AND PRIORITIES containing 2 initial questions about the roadmap of the installations and 9 questions to collect feedback on the following topics:
  - o Support provided / to be provided by the Dataverse software
  - o Support provided / to be provided by the Dataverse community / GDCC
  - o Sustainability and Governance
- PART 5: OTHER FEEDBACK asking the participants for any other feedback on the survey

All questions in the consent section were mandatory questions.

In addition to the questions in the consent section, the survey form contained altogether 43 questions. Of these, only Question 1 and Question 2 were mandatory. Most of the questions were closed-ended multiple choice questions including "Other" as one of the response options (Sauro, 2018). The "Other" option could be elaborated in an open-ended / free-text field following the closed-ended question. In all open-ended questions, there was a reminder to respondents not to add any personal information.

Under question 1, the survey participants were also informed that if they were planning to have a production Dataverse installation or were considering the Dataverse software for their current or future repository, they should answer the subsequent questions based on what they expect will or could be applying for their repository when/if they deploy the Dataverse software.

Also under Question 1, the participants were informed that unless indicated otherwise, the questions in the survey apply to (existing/planned/considered) production installations, not testing or staging installations.

The content of the questions was designed based on previous discussions in the community. An earlier version of the survey questions was discussed with several members of the community; see the *Acknowledgement* section above.

The final version of the questions was implemented in the survey tool Nettskjema (https://nettskjema.no/?lang=en).

Survey distribution

The Dataverse Community Survey 2022 was conducted between May 11th and June 7th, 2022. It was first distributed on May 11th, 2022, to Dataverse installations by sending an invitation to the contact email addresses which someone representing the installation added to a shared Google spreadsheet which is maintained by the Dataverse community. For installations lacking an email contact address in the spreadsheet and in cases where the provided email turned out not to be valid or in use anymore, an invitation was submitted to the contact form of the installation. In this original invitation, the submission deadline was set to May 31st, 2022. This invitation was sent to 78 existing Dataverse installations. In addition, the invitation was sent to 6 organizations that were planning to deploy or considering deploying the Dataverse for their repository.

A reminder about the survey was posted to the Google Group of the Dataverse Users Community (dataverse-community@googlegroups.com) on May 23rd, 2022.

Deadline extension until June 7th, 2022, was announced in the Google Group on June 3rd, 2022.

The participants were asked to submit their responses in the Nettskjema survey tool. The survey questions were also provided in a text document version; see the file "Dataverse_Community_Survey_2022{.docx,.pdf}" in the supporting dataset for the survey report (Conzett, 2022).

Data Collected

During the survey conducting period a total of 34 responses were collected using the Nettskjema survey tool. 32 of these responses were from existing Dataverse installations, whereas 2 responses were from organizations that are planning to install Dataverse. The 32 responses from existing Dataverse installations represent 41% of the 78 Dataverse installations which at survey start were registered in the overview provided on the Dataverse Project website (https://dataverse.org/).

The anonymized raw data as well as processed data for the analysis of most of the survey questions are available in the supporting dataset for the survey (Conzett, 2022). The supporting dataset also contains a detailed description of how the raw data was processed to support the analysis presented in this present report.

Where applicable, the names of the files containing the processed data that were used for the analysis of the survey results from a given question are added at under the heading "Supporting material" at the end of the presentation of that question in the *Results* section below. In addition to the processed data, the anonymized raw data in the file "DCS_2022_data{.tsv,.xlsx}" may also be consulted, but this file name is not added to the file list under "Supporting material" unless it was the only data file that was used for the analysis of the results from that question.

Results

PART 2: RELATION TO DATAVERSE PROJECT

Question 1. Status: Indicate the status of your Dataverse installation / your interest in the Dataverse community.

This mandatory, closed-ended, single-choice question resulted in the following distribution of the submitted responses across the four possible response options:
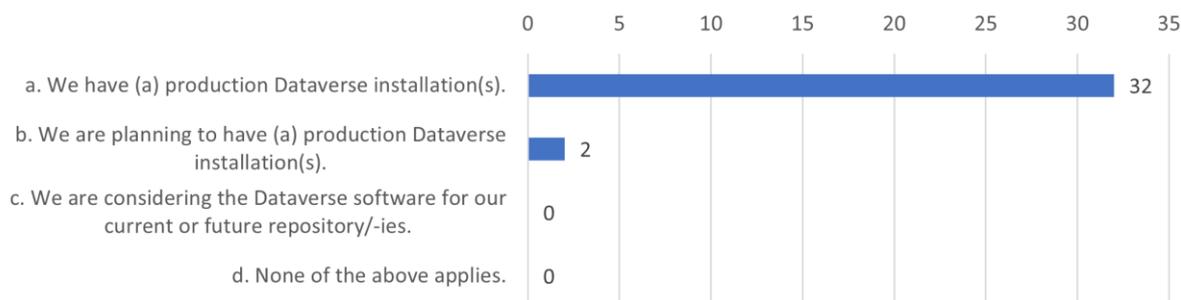

Figure 1: Results from Question 1. Status: Indicate the status of your Dataverse installation / your interest in the Dataverse community. The figure shows the number of submitted responses for each option.

32 of the respondents represent (a) production Dataverse installation(s), whereas two of the represented installations are still in planning.

```
### Supporting material:
DCS_2022_Q01_Status_data{.tsv,.xlsx}
DCS_2022_Q01_Status_chart.png
```

PART 3: INSTALLATION INFORMATION

Question 2. Installation Name: Enter the name of your Dataverse installation.

This mandatory, open-ended question resulted in the following 34 responses, sorted alphabetically:

- ADA Dataverse
- ASU Library Research Data Repository
- AUSSDA
- Borealis, the Canadian Dataverse Repository (formerly Scholars Portal Dataverse installation)
- CedapDados
- CIMMYT Research Data & Software Repository Network
- Cirad Dataverse
- Concorda
- CORA.Repositori de Dades de Recerca

- DaRUS
- Data SciELO
- data.sciencespo
- Dataverse[1]
- dataverse.ucla.edu
- DataverseNL
- DataverseNO
- DOREL
- e-cienciaDatos
- Edmond
- George Mason University Dataverse
- Göttingen Research Online / Data
- Harvard Dataverse
- IBICT Cariniana Dataverse Network
- IIT Dataverse
- Indata
- Jülich DATA
- KEEN Dataverse
- KU Leuven RDR
- LiDA Dataverse repository
- Qualitative Data Repository
- Repositori Ilmiah Nasional (RIN)
- Repositorio de datos de investigación de la Universidad del Rosario
- UVa Dataverse: LibraData
- VTTI

Question 3. Installation URL: Enter the URL of your Dataverse installation.

For the responses to this optional, open-ended question, see the column "Q03_Installation_URL" in the anonymized raw data file "DCS_2022_data{.tsv,.xlsx}".

Question 4. re3data: Enter the DOI of the re3data record of your Dataverse installation.

This optional, open-ended question resulted in the five types of repsonses summarized in Figure 2.

---

[1] This answer is from a participant representing a Dataverse installation that is still under planning, thus the name "Dataverse" was used as a placeholder. After the survey answering deadline, the name of this installation was announced to be "DeiC Dataverse"; cf. https://deic.dk/en/node/2147.
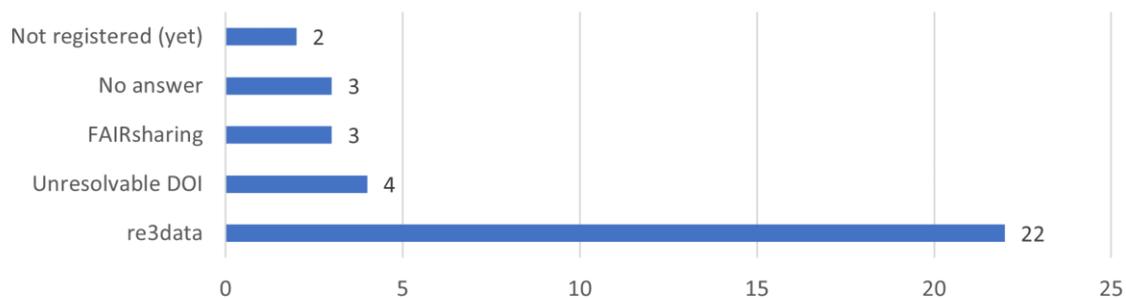
Figure 2: Results from Question 4. re3data: Enter the DOI of the re3data (https://www.re3data.org/) record of your Dataverse installation (e.g., http://doi.org/10.17616/R3C880). The figure shows the number of submitted responses for each type of unique response.

With 22 responses, the majority (65%) of respondents entered a DOI pointing to the re3data record of their Dataverse installation. Four respondents entered a non-resolvable DOI. Three respondents entered a DOI pointing to the FAIRsharing record of their Dataverse installation (cf. https://fairsharing.org/). A future survey should probably include a question about FAIRsharing. Another three respondents left the field empty. Two respondents answered that they had not (yet) registered their Dataverse installation in re3data.

```
### Supporting material:
DCS_2022_Q04_re3data_data{.tsv,.xlsx}
DCS_2022_Q04_re3data_type_of_response_data{.tsv,.xlsx}
DCS_2022_Q04_re3data_chart.png
```

Question 5. Country: Enter the name of the country where your Dataverse installation is based.

In this optional, closed-ended, single-choice question, the respondents could choose from a dropdown list containing all ISO 3166-1-alpha-2 English country names and code elements, as provided by DataHub (n.d.), e.g., Afghanistan (AF), Albania (AL), Algeria (DZ), etc.

Of the 34 Dataverse installations represented in the submitted answers, 7 were from the United States (US), 5 from Germany (DE), 3 from Brazil (BR), 3 from France (FR), 2 from Spain (ES), and 1 from each of the following countries: Australia (AU), Austria (AT), Belgium (BE), Canada (CA), Colombia (CO), Denmark (DK), Ecuador (EC), Hungary (HU), Indonesia (ID), Italy (IT), Lithuania (LT), Mexico (MX), Netherlands (NL), Norway (NO).

These countries are marked with different colours on the map in Figure 3, where each colour indicates the number of represented Dataverse installations in the survey.
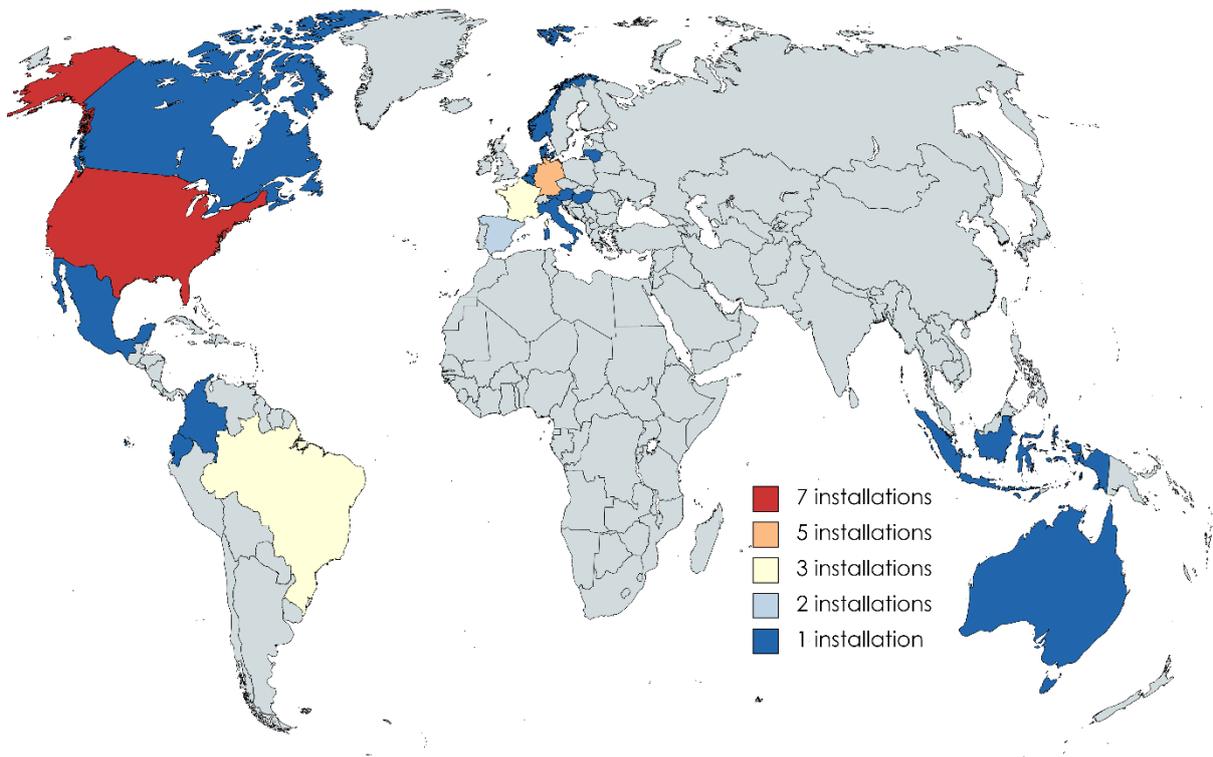
Figure 3: World map indicating the countries of Dataverse installations participating in the survey. The different colours indicate the numbers of Dataverse installations represented in the survey. The map was created with mapchart.net and is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

```
### Supporting material:
DCS_2022_Q05_Country_data{.tsv,.xlsx}
```

Question 6. Continent: Enter the name of the continent where your Dataverse installation is based.

In this optional, closed-ended, single-choice question, the respondents could choose from a dropdown list containing the following names of continent: Asia, Australia, Europe, North America, South America.

The distribution of received responses is as follows: Europe: 18, North America: 9, South America: 5, Asia: 1, Australia: 1. Thus more than half (53%) of the submitted responses were from Dataverse installations based in Europe.

```
### Supporting material:
DCS_2022_Q06_Continent_data{.tsv,.xlsx}
```

Question 7. Certification: Has your Dataverse installation obtained any repository (or similar) certification?

In this optional, closed-ended, multiple-choice question, the respondents were asked to choose between the following options: "Currently certified", "Planning to (re)apply", "No".

Figure 4 shows the number of responses for "Currently certified", "Planning to (re)apply", whereas the option "No" as well as empty answers are not included in the figure.
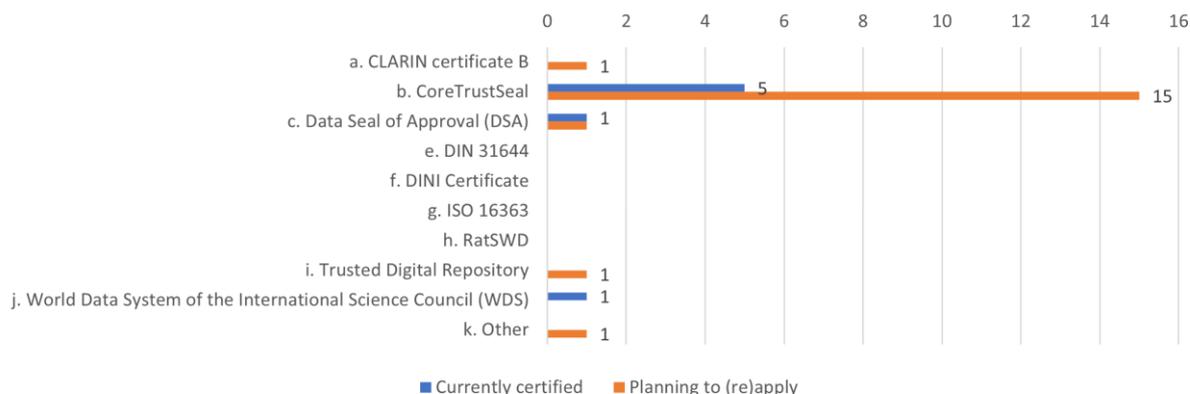

Figure 4: Results from Question 7. Certification: Has your Dataverse installation obtained any repository (or similar) certification? The figure shows the number of submitted responses for each option.

CoreTrustSeal is the most popular certification program among the Dataverse installations participating in the survey. More than have of the Dataverse installations are either CoreTrustSeal-certified (5 installations) or are planning to (re)apply for CoreTrustSeal certification (15 installations). In one case, the certification only applies to one repository contained in the installation. In addition, one respondent did not select the response alternative "Planning to (re)apply", but answered in the "Other" field that they are planning to apply for the CoreTrustSeal. One respondent selected the alternative "Planning to (re)apply" for the certification option "k. Other" and specified the option as "List of certified repositories in Hungary".

```
### Supporting material:
DCS_2022_Q07_Certification_data{.tsv,.xlsx}
DCS_2022_Q07_Certification_chart.png
```

Installation Type

Question 8. Installation Archetype: Select the archetype of your Dataverse installation.

This optional, closed-ended, multiple-choice question asks about the archetype of the Dataverse installation according to the classification which resulted from a study of Dataverse installations carried out from 2019 to 2021 by several members of the Dataverse community (Schlatter & Ji, 2021).

The number of responses for each option is summarized in Figure 5. More than half of the 34 Dataverse installations participating in the survey have an institutional focus. Eight of the installations classify themselves as domain specific, whereas five have a global focus. Five installations selected the option "Other". Two of them commented that they run a national repository, whereas one respondent classified their installation as

consortium focused. For the remaining two comments, see the anonymized raw data file
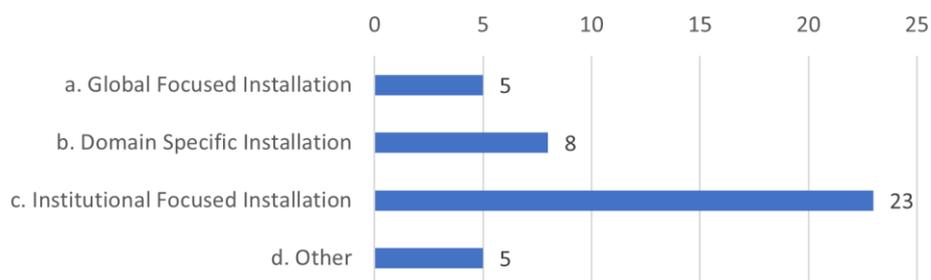"DCS_2022_data{.tsv,.xlsx}".

Figure 5: Results from Question 8. Installation Archetype: Select the archetype of your Dataverse installation. The figure shows the number of submitted responses for each selected response option.

In Figure 6, the percentages of responses to Question 8 of the Dataverse Community Survey 2022 are compared with the percentages of installations reported by Schlatter and Ji (Schlatter & Ji, 2021). There are only slight differences in the share of Installation Archetypes between the two studies.
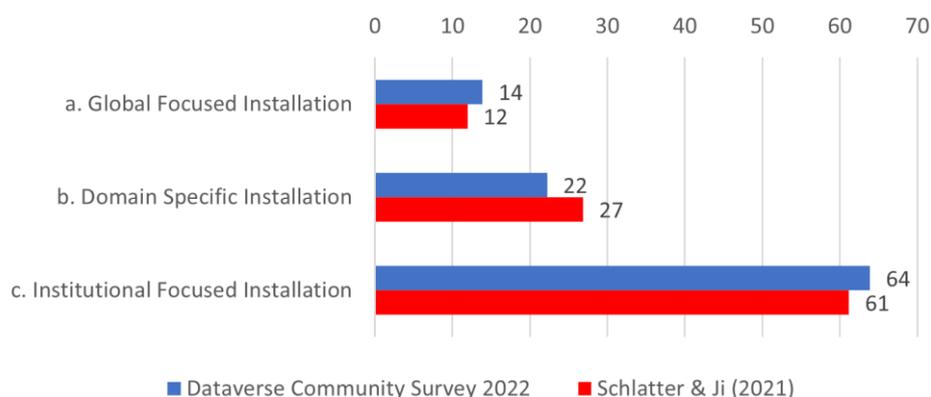

Figure 6: Results from Question 8 compared to the results reported in Schlatter & Ji (2021). The figure shows the percentages of responses in the Dataverse Community Survey 2022, and the percentages of installations reported in Schlatter & Ji (2021). The response option "Other" is not included in the percentages from the Dataverse Community Survey 2022.

```
### Supporting material:
DCS_2022_Q08_Installation_Archetype_data{.tsv,.xlsx}
DCS_2022_Q08_Installation_Archetype_chart.png
DCS_2022_Q08_Installation_Archetype_compared_to_Schlatter_and_Ji_202
  1_data{.tsv,.xlsx}
DCS_2022_Q08_Installation_Archetype_compared_to_Schlatter_and_Ji_202
  1_chart.png
```

Question 9. Domain: If your installation has a domain specific focus, please specify the high-level discipline(s) to which the domain(s) belong(s).

In this optional, closed-ended, multiple-choice question, participants representing a Dataverse installation with domain specific focus (cf. Question 8) were asked to select the

high-level discipline(s) to which the domain(s) their installation serves belong to. The high-level disciplines used in this survey correspond to the six first digit level / major fields described in the 2007 edition of the OECD Frascati classification of science and technology (FOS) (Working Party of National Experts on Science and Technology Indicators, 2007). In addition, the response option "Other" was added and could be specified in a separate field:

- a. Natural Sciences
- b. Engineering and Technology
- c. Medical and Health Sciences
- d. Agricultural Sciences
- e. Social Sciences
- f. Humanities
- g. Other

Of the submitted responses, several of these installations do actually not have a domain specific focus (see Question 8) and were thus not supposed to answer this question. The most typical combination of disciplines that was selected is abcdef, meaning that the installation serves all disciplines. Also, five of the 22 installations that responded to this question selected the option "Other" as their only choice. Three of them commented that their installation is multidisciplinary. A fourth one specified the installation as a national one. In future surveys, this question should probably be designed as a conditional question which is only displayed for domain-focused installations.

In the following, only responses from installations with a domain specific focus are discussed.

Figure 7 summarizes the numbers of responses from domain-focused installations for the different selected high-level disciplines and combinations of those. The names of discipline are abbreviated using only the initial letter in the options listed above, e.g., "a" for "Natural Sciences", etc.

Only eight of the 34 Dataverse installations participating in the survey have a domain-specific focus. Four of them serve a combination of disciplines (ab, ad, cef, ef), and three of them focus on one discipline (d, e). Among all these domain-specific installations, Social Sciences is with a total of four installations the most common discipline being served, which does not come as a surprise as the Dataverse software has its origin in quantitative social sciences.
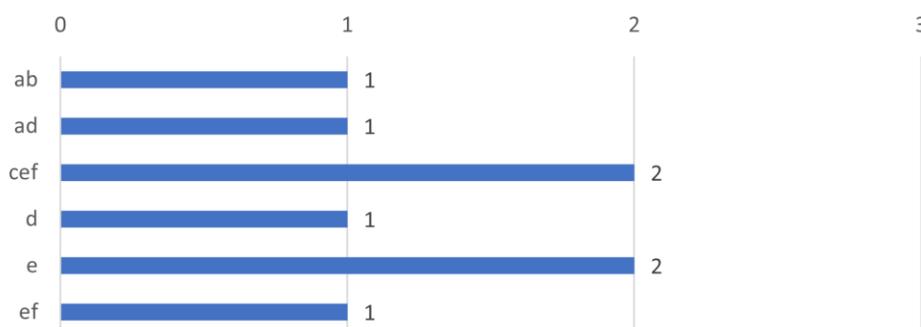
Figure 7: Results from Question 9. Domain: If your installation has a domain specific focus, please specify the high-level discipline(s) to which the domain(s) belong(s). Abbreviations: a = Natural Sciences; b = Engineering and Technology; c = Medical and Health Sciences; d = Agricultural Sciences; e = Social Sciences; f = Humanities. The figure shows the number of submitted responses from installations with domain-specific focus for each selected discipline or combination of disciplines.

```
### Supporting material:
DCS_2022_Q09_Domain_data{.tsv,.xlsx}
DCS_2022_Q09_Domain_chart.png
DCS_2022_Q09_Domain_type_of_response_data{.tsv,.xlsx}
```

Organization

Questions 10 to 13. Name, ROR ID, Multi-Institutional, Number of Members

Questions 10 to 13 are about the organization responsible for the operation of the Dataverse installation and specifically ask for the following information:

- Question 10. Organization Name: Enter the name of the organization which is responsible for the operation of your Dataverse installation.
- Question 11. ROR ID: If applicable, enter the ROR ID of the organization (cf. https://ror.org/; e.g, https://ror.org/03vek6s52).
- Question 12. Multi-Institutional: If your Dataverse installation is organized as a consortium or similar multi-institutional organization, enter the name of the multi-institutional organization.
- Question 13. Number of Members: If your Dataverse installation is organized as a consortium or similar multi-institutional organization, how many institutions are currently participating?

For the responses to these questions, see columns "Q10_Organization_Name", "Q11_ROR_ID", "Q12_Multi-Institutional", and "Q13_Number_of_Members" in the anonymized raw data file "DCS_2022_data{.tsv,.xlsx}".

There are twelve multi-institutional Dataverse installations participating in the survey. Figure 8 gives an overview of the number of institutions they serve.
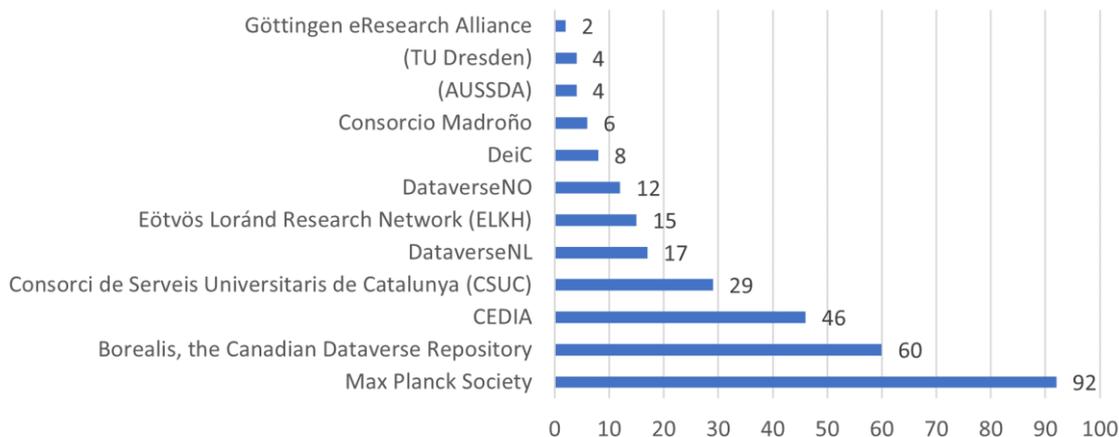
Figure 8: Results from Question 12 and Question 13. The figure shows the number of members of twelve multi-institutional Dataverse installations.

These data are slightly processed; see the file "0_README.txt" in the supporting dataset (Conzett, 2022) for more information.

Together, these twelve installations serve 295 institutions.

```
### Supporting material:
DCS_2022_Q12_Multi-institutional_data{.tsv,.xlsx}
DCS_2022_Q12_Multi-institutional_chart.png
```

Deployment

The aim of this section of the survey is to get an overview of different aspects of the software deployment used at the different Dataverse installations.

Question 14. Software Distribution: Which distribution of the Dataverse software does your production installation use?

In this optional closed-ended single choice question, the respondents were asked to choose between the options "a. Main distribution", "b. Fork", and "c. Other". Figure 9 shows the number of submitted responses for each selected response option.
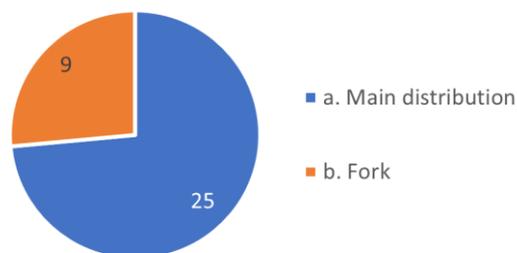

Figure 9: Results from Question 14. Software Distribution: Which distribution of the Dataverse software does your production installation use? The figure shows the number of submitted responses for each selected response option.

Of the 34 Dataverse installations participating in the survey, 25 (74%) use the main distributions of the Dataverse software, whereas 9 (26%) use a fork. One respondent specified in the "Other" field that they made some local customizations. This might also be true for other installations using the main distribution of the Dataverse software. Another respondent added that they use a fork, but that their aim is to use the main distribution.

```
### Supporting material:
DCS_2022_Q14_Software_Distribution_data{.tsv,.xlsx}
DCS_2022_Q14_Software_Distribution_chart.png
```

Question 15. Software Version: Which version of the main distribution of the Dataverse software is your production installation based on?

In this optional closed-ended single choice question, respondents were asked to choose a version number from a pull-down menu. The version numbers were retrieved from the release overview on the Dataverse GitHub website (*IQSS/Dataverse/Releases*, n.d.), which includes releases back to version 4.0.

Figure 10 shows the number of submitted responses for each selected response option. A vast majority of the installations participating in the survey were based on version 5.10.1, which at that time was the latest released version of the main distribution of the Dataverse software.
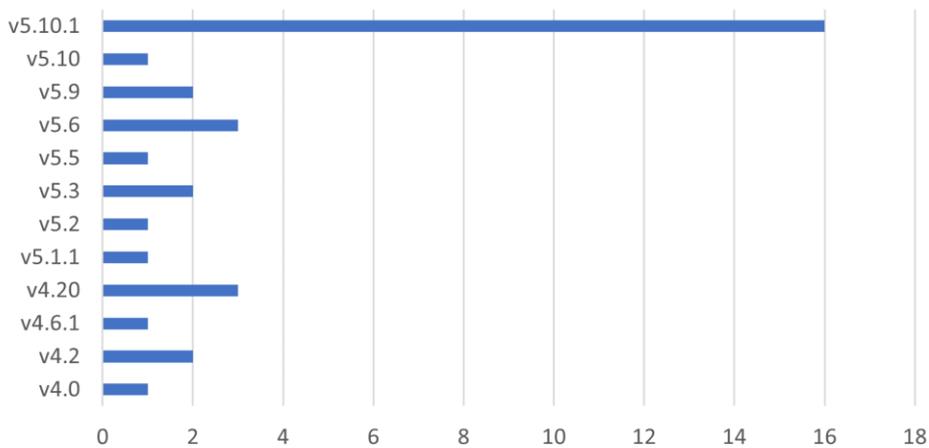

Figure 10: Results from Question 15. Software Version: Which version of the main distribution of the Dataverse software is your production installation based on? The figure shows the number of submitted responses for each selected response option.

```
### Supporting material:
DCS_2022_Q15_Software_Version_data{.tsv,.xlsx}
DCS_2022_Q15_Software_Version_chart.png
```

Questions 16 to 20 are about different aspects of software deployment. They are all optional closed-ended multiple choice questions. In each question, the respondents were

asked to indicate which given response options that apply to their current, planned or considered deployment of the Dataverse software. In some cases, some of the installations not yet deployed in production (see Question 1) opted for "Current deployment" options. They probably overlooked the remark in Question 1 that the questions in this survey apply to (existing/planned/considered) *production* installations, not testing or staging installations.

Question 16. Software Deployment Location: Indicate where the Dataverse software is deployed in your production installation.

In Question 16, respondents were asked to indicate where the Dataverse software is deployed in their production environment. Figure 11 shows the number of submitted responses for each selected response option.

Focusing on the situation for current production deployment, the chart tells us that 23 (67%) of the 34 Dataverse installations participating in the survey are deployed on (a) local server(s), 6 are using Amazon Web Services (AWS), and 5 are using another cloud service. Four of the five other cloud services used for current deployment were specified in the *Other* field as follows:

- Aptum Technologies private cloud environment
- internal computing center
- Ontario Library Research Cloud (SWIFT-S3 setup), locally managed with 5 nodes across Canada
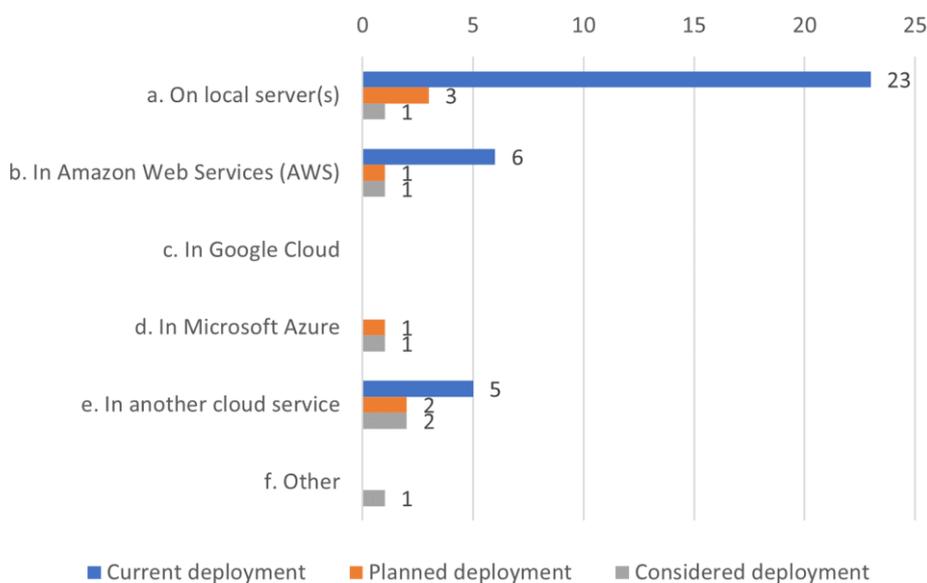- Own Data Center cloud manage by OpenNebula

Figure 11: Results from Question 16. Software Deployment Location: Indicate where the Dataverse software is deployed in your production installation. The figure shows the number of submitted responses for each selected response option.

For more details about deployment locations for planned and considered deployments, see the anonymized raw data file "DCS_2022_data{.tsv,.xlsx}".

```
### Supporting material:
DCS_2022_Q16_Software_Deployment_Location_data{.tsv,.xlsx}
DCS_2022_Q16_Software_Deployment_Location_chart.png
```

Question 17. Software Deployment Method: Indicate how the Dataverse software is deployed in your production installation.

In Question 17, respondents indicated the method they use to deploy their Dataverse software in their production environment. Figure 12 shows the number of submitted responses for each selected response option.
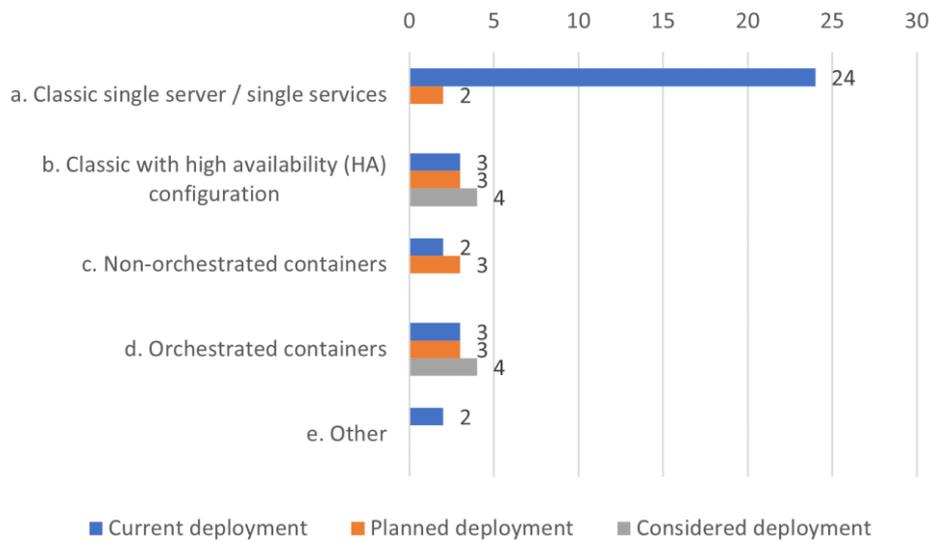

Figure 12: Results from Question 17. Software Deployment Method: Indicate how the Dataverse software is deployed in your production installation. The figure shows the number of submitted responses for each selected response option.

Concentrating again on current deployment, the majority (24 or 71%) of the participating 34 installations are using a classic single server / single services method; 3 are using a classic method with high availability (HA) configuration; 2 are using non-orchestrated containers; 3 are using orchestrated containers. 10 installations are either planning to use or considering using a container-based method to deploy the Dataverse software.

For other details, see the anonymized raw data file "DCS_2022_data{.tsv,.xlsx}".

```
### Supporting material:
DCS_2022_Q17_Software_Deployment_Method_data{.tsv,.xlsx}
DCS_2022_Q17_Software_Deployment_Method_chart.png
```

Question 18. Software Deployment Tools: Indicate the tool(s) you use to deploy the Dataverse software in your production installation.

In Question 18, respondents specified which (if any) tool(s) they use to deploy the Dataverse software in their production environment. Figure 13 shows the number of submitted responses for each selected response option.
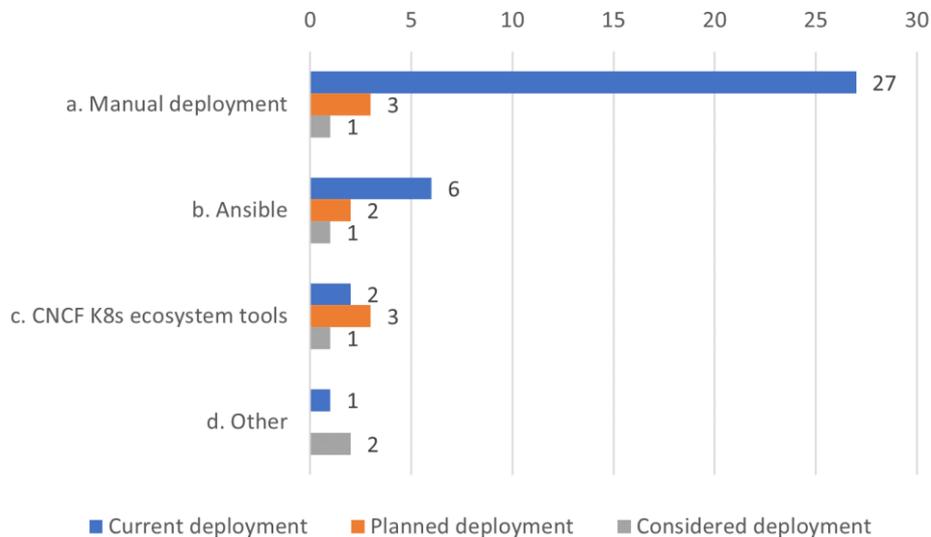

Figure 13: Results from Question 18. Software Deployment Tools: Indicate the tool(s) you use to deploy the Dataverse software in your production installation. The figure shows the number of submitted responses for each selected response option.

27 or almost 80% of the 34 installations participating in the survey are using manual deployment. 9 installations are using a deployment tool; 6 are using Ansible, 2 are using CNCF K8s ecosystem tools; 1 us using another tool, specified as Terraform in the *Other* field.

Together with the previous two questions, the following pattern emerges: Installations on local server(s) using classic deployment methods are deployed manually, whereas installations located in a cloud infrastructure are deployed with containers using one or more deployment tools.

For other details, see the anonymized raw data file "DCS_2022_data{.tsv,.xlsx}".

```
### Supporting material:
DCS_2022_Q18_Software_Deployment_Tools_data{.tsv,.xlsx}
DCS_2022_Q18_Software_Deployment_Tools_chart.png
```

Question 19. File Storage Type: Indicate the type of file storage used for the research data files stored in your Dataverse production installation.

The next two questions are about the file storage used for the Dataverse production installation. Question 19 asks about the type of file storage.

Figure 14 shows the number of submitted responses for each selected response option.
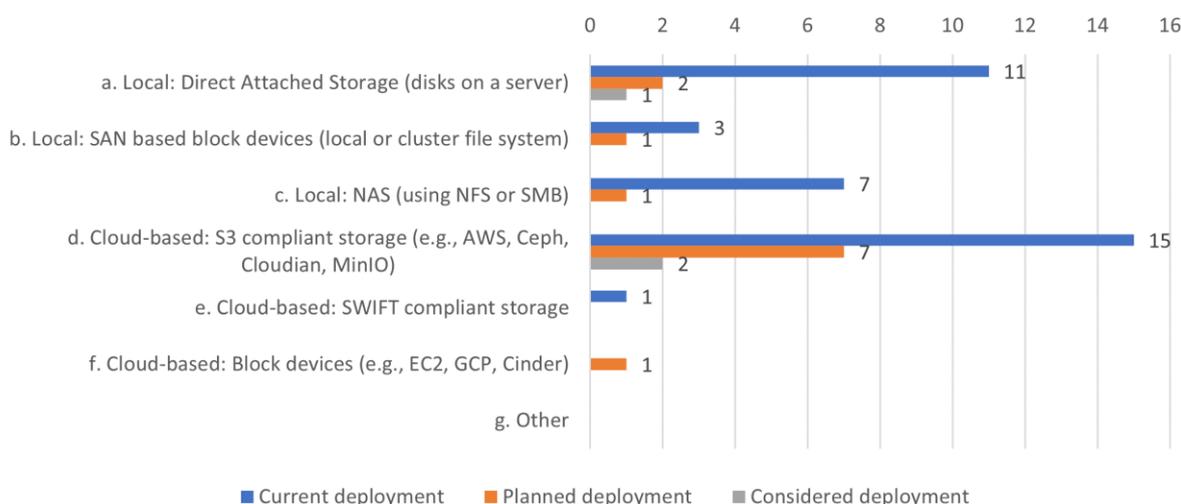

Figure 14: Results from Question 19. File Storage Type: Indicate the type of file storage used for the research data files stored in your Dataverse production installation. The figure shows the number of submitted responses for each selected response option.

21 (62%) of the 34 installations participating in the survey use local file storage in the current deployment of their Dataverse installations, whereas 16 (47%) installations use cloud-based file storage, meaning that 3 installations are using both local and cloud-based file storage. This means that some installations combine cloud-based storage with a local deployment of the application; see Question 16.

Of the 21 installations using local file storage in their current deployment, more than half (11) use Direct Attached Storage (disks on a server); 3 use SAN based block devices (local or cluster file system); 7 use SAN based block devices (local or cluster file system).

Of the 16 installations using cloud-based file storage in their current deployment, all but one use S3 compliant storage (e.g., AWS, Ceph, Cloudian, MinIO); one uses SWIFT compliant storage. One installation is planning to use block devices (e.g., EC2, GCP, Cinder) for their cloud-based file storage.

```
### Supporting material:
DCS_2022_Q19_File_Storage_Type_data{.tsv,.xlsx}
DCS_2022_Q19_File_Storage_Type_chart.png
```

Question 20. File Storage Provider: Indicate the provider of the research data file storage of your Dataverse production installation.

Following up on the previous question, in Question 20 respondents were asked to indicate the provider of the data file storage used in their Dataverse production installation.

Figure 15 shows the number of submitted responses for each selected response option.
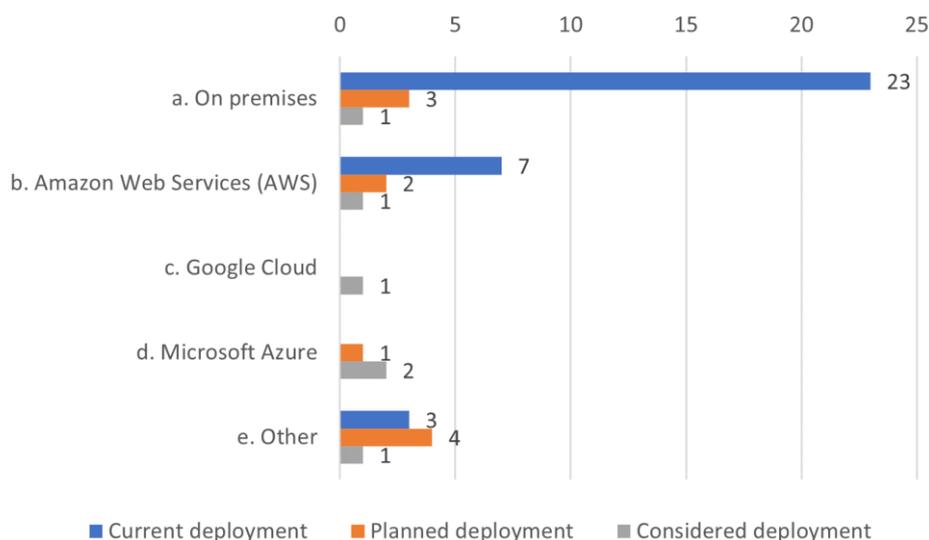
Figure 15: Results from Question 20. File Storage Provider: Indicate the provider of the research data file storage of your Dataverse production installation. The figure shows the number of submitted responses for each selected response option.

Keeping our focus on current deployment, the figure reveals that 23 (68%) of the 34 Dataverse installations participating in the survey use file storage on premises; 7 of them use cloud-based file storage provided by Amazon Web Services (AWS). 3 installations use another file storage provider for their current deployment, specified as follows in the *Other* field:

- Aptum Technologies for current deployment
- Ceph
- Ontario Library Research Cloud (distributed, local copy stored at UofT)

For other details, see the anonymized raw data file "DCS_2022_data{.tsv,.xlsx}".

```
### Supporting material:
DCS_2022_Q20_File_Storage_Provider_data{.tsv,.xlsx}
DCS_2022_Q20_File_Storage_Provider_chart.png
```

Configurable/Optional Features

This section of the survey aims at mapping which configurable and optional features Dataverse production installations are currently using, planning to use, or considering using. Questions 21 to 29 are all optional closed-ended single choice questions, where respondents for each feature could select between the following response options: "Using", "Planning", "Considering", "Not using", "Don't know". In addition, respondents were given the possibility to elaborate in an open-ended / free-text field. For reasons of readability, only the first three options, i.e., "Using", "Planning", "Considering", are

illustrated in the figures in the presentation below. For other details, the reader is referred to the supporting material.

For several of the features, some respondents commented that they were unsure about the usage and/or that the feature might be used without the repository management being aware of this.

Question 21. Configurable/Optional Features: Getting Data In

Question 21 is about configurable/optional features or tools that can be used to get data into a dataset created in a Dataverse installation. Figure 16 shows the number of submitted responses for three selected response options.
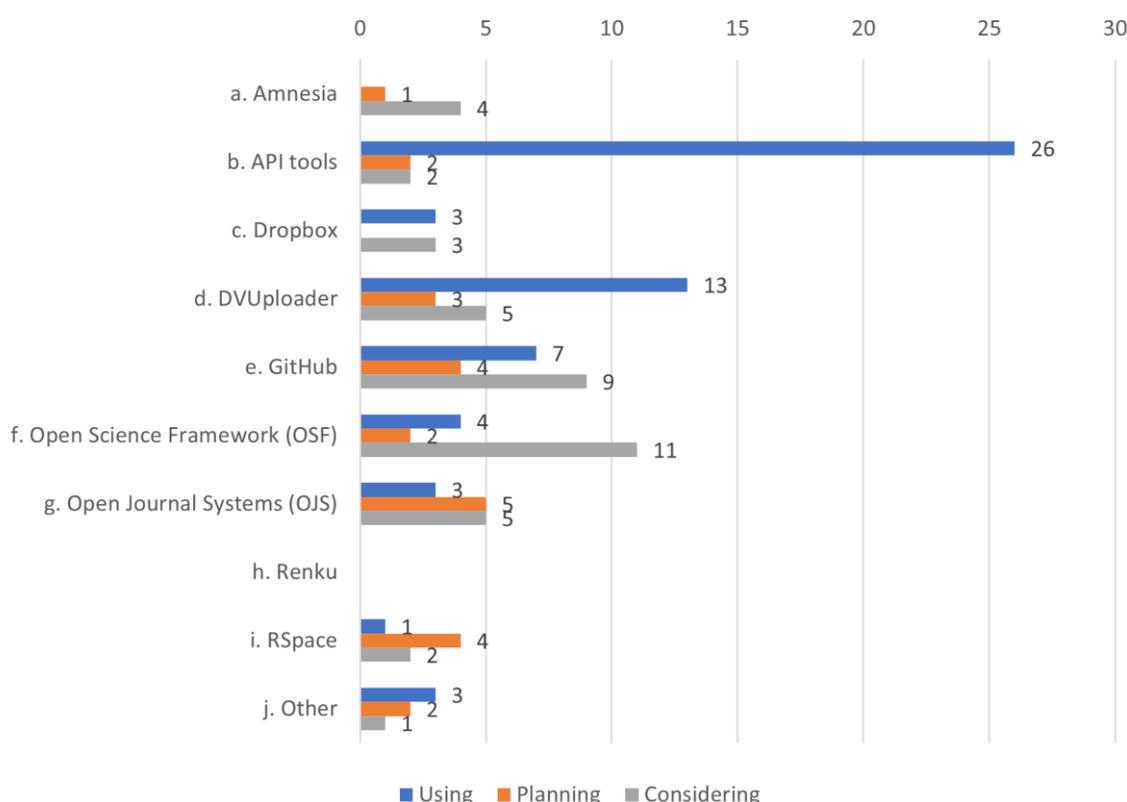

Figure 16: Results from Question 21. Configurable/Optional Features: Getting Data In. The figure shows the number of submitted responses for three selected response options.

By far the most used feature or tool to get data into a dataset are API tools. 26 of the 34 Dataverse installations participating in the survey use API tool for this purpose. Half as many installations (13) use the DVUploader, and 7 installations use GitHub to get data in. Among the other features and tools, Open Science Framework (OSF) and Open Journal Systems (OJS) are either planned or considered by 10 or more installations (OSF: 13; OJS: 10).

In the *Other* field, the following features and tools were specified:

- Using: GeoMapper, Matomo, Python interface
- Planning: iRODS, Gitlab, Globus integrated tool
- Considering: Gitlab

For other features and tools, see Figure 16 and the supporting material.

```
### Supporting material:
DCS_2022_Q21_Configurable_Optional_Features_Getting_Data_In_data{.ts
  v,.xlsx}
DCS_2022_Q21_Configurable_Optional_Features_Getting_Data_In_chart.pn
  g
```

Question 22. Configurable/Optional Features: API Client Libraries

In Question 22, respondents were asked to indicate to what extent their Dataverse installations use available API client libraries. Figure 17 shows the number of submitted responses for three selected response options.
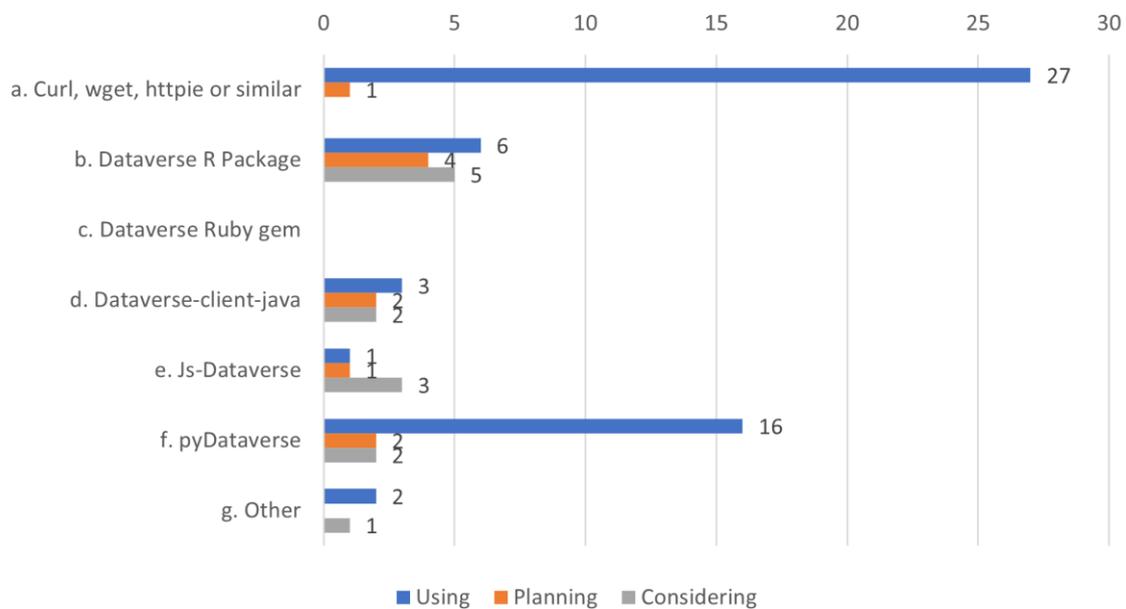

Figure 17: Results from Question 22. Configurable/Optional Features: API Client Libraries. The figure shows the number of submitted responses for three selected response options.

The three most used API client libraries are Curl, wget, httpie or similar (used by 27 of the 34 Dataverse installations participating in the survey), pyDataverse (used by 16 installations), and the Dataverse R Package (used by 6 installations). The latter is also planned to be or considered being used by 9 installations.

For other features and tools, see Figure 17 and the supporting material.

Question 23. Configurable/Optional Features: Embedding Data on Websites

In Question 23, respondents were asked to indicate to what extent their Dataverse installations use features/tools to embed published datasets on websites. Figure 18 shows the number of submitted responses for three selected response options.
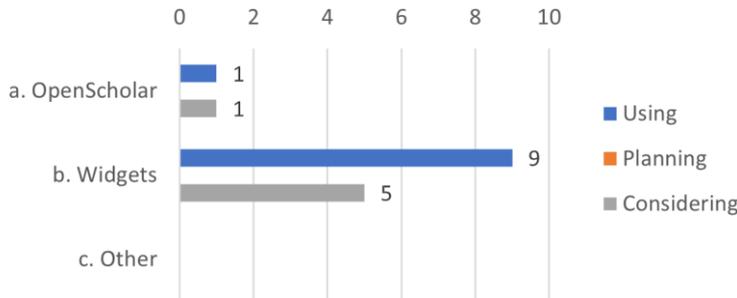

Figure 18: Results from Question 23. Configurable/Optional Features: Embedding Data on Websites. The figure shows the number of submitted responses for three selected response options.

Widgets are the only features/tools in this category that are more widely used or considered being used at Dataverse installations to embed data on websites. Of the 34 Dataverse installations participating in the survey, 9 use widgets, whereas 5 consider using them.

For other features and tools, see Figure 18 and the supporting material.

```
### Supporting material:
DCS_2022_Q23_Configurable_Optional_Features_Embedding_Data_on_Websit
  es_data{.tsv,.xlsx}
DCS_2022_Q23_Configurable_Optional_Features_Embedding_Data_on_Websit
  es_chart.png
```

Question 24. Configurable/Optional Features: Analysis and Computation

In Question 24, respondents were asked to indicate to what extent their Dataverse installations use features/tools for analysis and computation of published datasets.

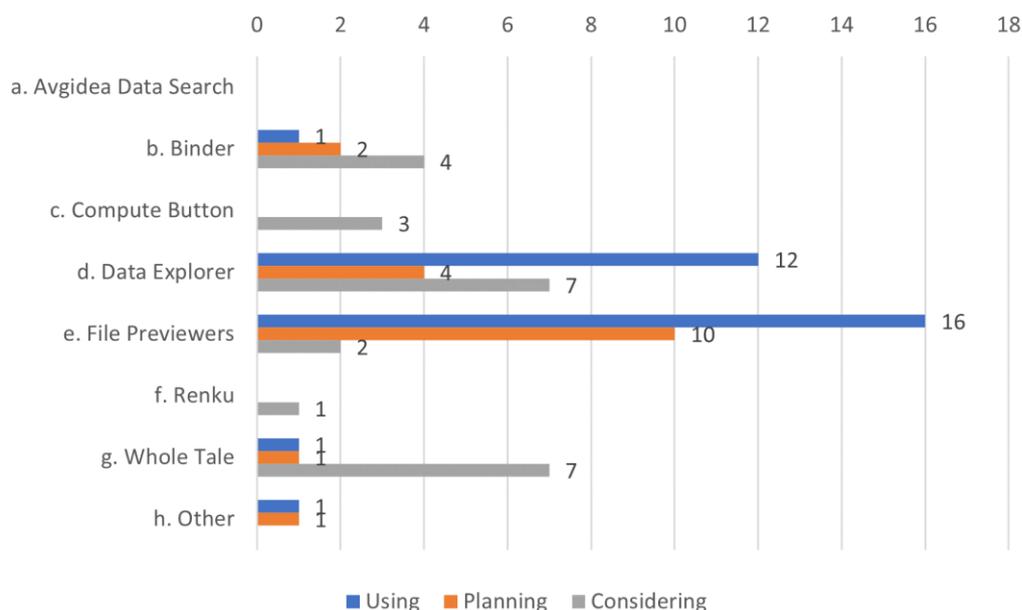Figure 19 shows the number of submitted responses for three selected response options.

Figure 19: Results from Question 24. Configurable/Optional Features: Analysis and Computation. The figure shows the number of submitted responses for three selected response options.

File previewers and the Data Explorer are the two features/tools for data analysis and computation most widely used at the Dataverse installations participating in the survey. 16 installations already use, 10 installations are planning to use, and 2 installations are considering using file previewers. The corresponding numbers for the Data Explorer are 12 (Using); 4 (Planning); 7 (Considering).

The following features/tools were mentioned in the *Other* field:

- Using: Two Ravens
- Planning: OpenTOSCA, VipLab

For other features and tools, see Figure 19 and the supporting material.

```
### Supporting material:
DCS_2022_Q24_Configurable_Optional_Features_Analysis_and_Computation
  _data{.tsv,.xlsx}
DCS_2022_Q24_Configurable_Optional_Features_Analysis_and_Computation
  _chart.png
```

Question 25. Configurable/Optional Features: Discoverability

In Question 25, respondents were asked to indicate to what extent their Dataverse installations use features/tools to support discoverability of published datasets.

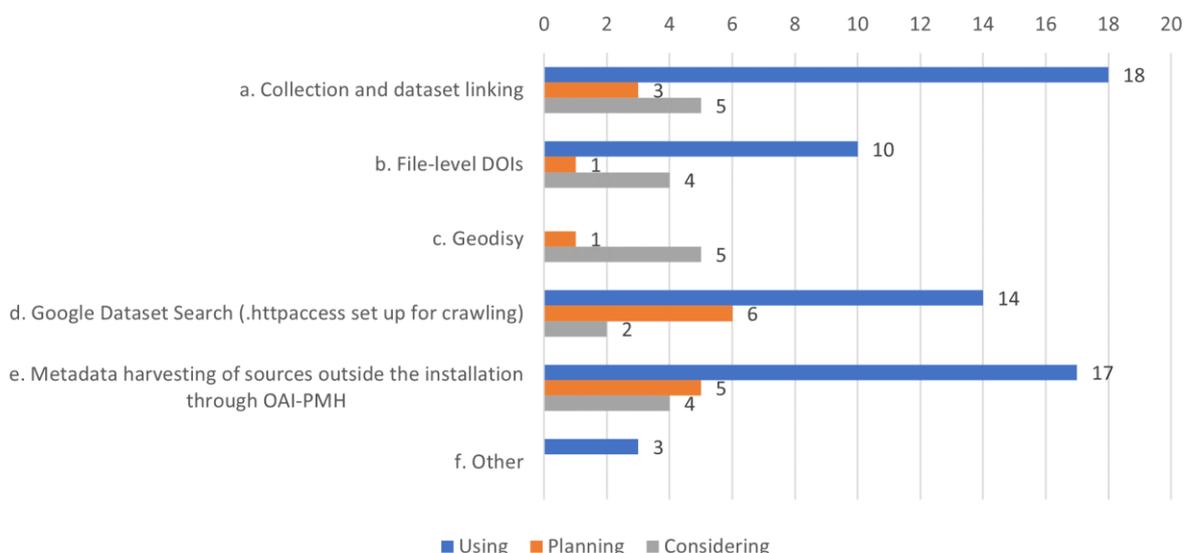Figure 20 shows the number of submitted responses for three selected response options.

Figure 20: Results from Question 25. Configurable/Optional Features: Discoverability. The figure shows the number of submitted responses for three selected response options.

With the exception of Geodisy, which is planned to be or considered being used by 6 of the 34 installations participating in the survey, all the other listed features/tools for supporting discoverability are either used, planned to be or considered being used by a considerable share of the installations, indicated with percentages in the list below:

- b. File-level DOIs: 44%
- d. Google Dataset Search (.httpaccess set up for crawling): 65%
- a. Collection and dataset linking: 76%
- e. Metadata harvesting of sources outside the installation through OAI-PMH: 76%

Option e is about using the Dataverse software to harvest metadata from outside the Dataverse installation, e.g., from other Dataverse installations or other repositories, thus using the Dataverse software as a harvesting client. However, one of the comments in the *Other* field may indicate that this feature by some respondents might have been interpreted as being about using the Dataverse software to provide metadata through OAI-PMH to discovery services, thus using the software as a harvesting server.

The following to features were added in the *Other* field as examples of other features being used to support discoverability:

- File-level Handles
- Solr index push to local Library discovery system

For other features and tools, see Figure 20 and the supporting material.

```
### Supporting material:
DCS_2022_Q25_Configurable_Optional_Features_Discoverability_data{.ts
  v,.xlsx}
```

```
DCS_2022_Q25_Configurable_Optional_Features_Discoverability_chart.pn
  g
```

Question 26. Configurable/Optional Features: Curation and Preservation

In Question 26, respondents were asked to indicate to what extent their Dataverse installations use features/tools to support curation and preservation of deposited data. Figure 21 shows the number of submitted responses for three selected response options.
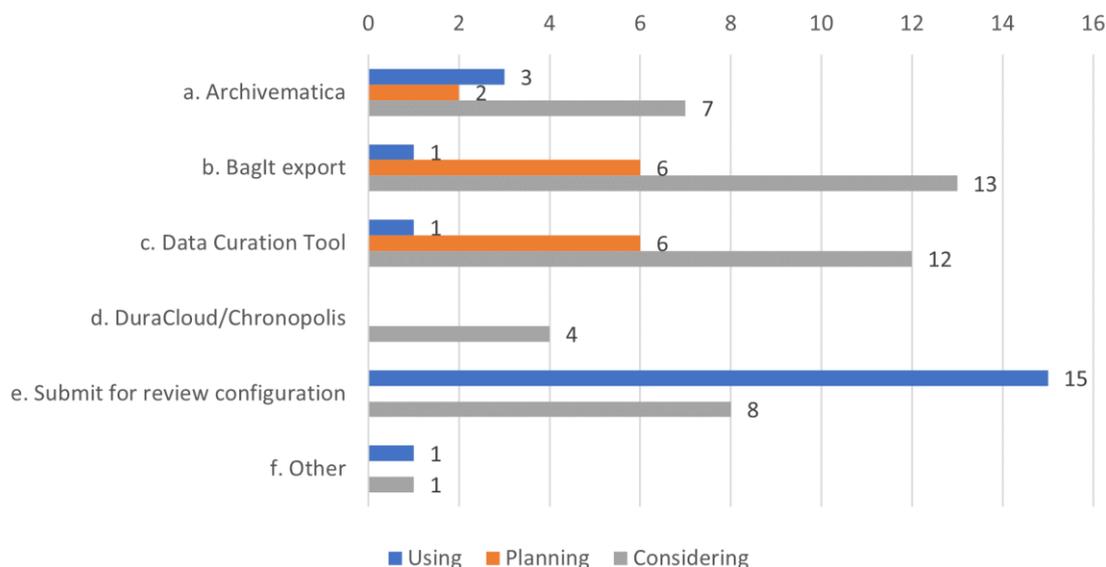

Figure 21: Results from Question 26. Configurable/Optional Features: Curation and Preservation. The figure shows the number of submitted responses for three selected response options.

Of the 34 Dataverse installations participating in the survey, 15 (44%) use the submit for review workflow provided by the software, and 8 installations consider using this feature. Although currently only used by one installation, both BagIt export and the Data Curation Tool are planned to be or considered being used by 19 installations (BagIt export) and 18 installations (Data Curation Tool).

For other features and tools, see Figure 21 and the supporting material.

```
### Supporting material:
DCS_2022_Q26_Configurable_Optional_Features_Curation_and_Preservatio
  n_data{.tsv,.xlsx}
DCS_2022_Q26_Configurable_Optional_Features_Curation_and_Preservatio
  n_chart.png
```

Question 27. Configurable/Optional Features: File Management

In Question 27, respondents were asked to indicate to what extent their Dataverse installations use features/tools to support the management of files deposited in or to be deposited in datasets.

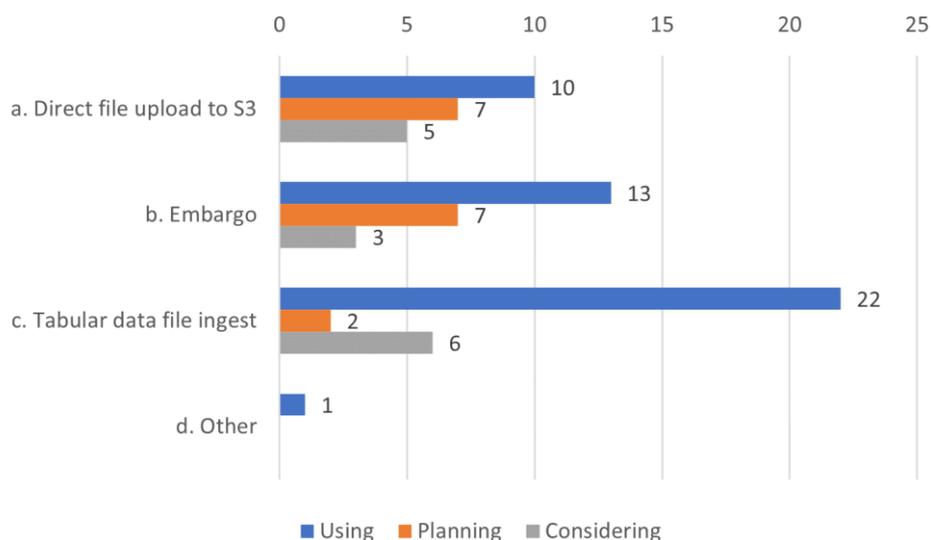Figure 22 shows the number of submitted responses for three selected response options.

Figure 22: Results from Question 27. Configurable/Optional Features: File Management. The figure shows the number of submitted responses for three selected response options.

Tabular data file ingest is used by 22 (65%) installations and thus the most commonly used file management features among the 34 Dataverse installations participating in the survey. 8 installations are planning to or considering using this feature. The automated embargo feature was introduced with version 5.8 in November 2021. Despite its recent introduction, the embargo feature is already used by 13 of the installations participating in the survey, and 10 installations are planning to or considering using the feature. Direct file upload to S3(-compliant) storage is used by 10 installations and planned to be or considered being used by 12 installations.

Wasabi file management backup is mentioned in the *Other* field as another file management feature used by a Dataverse installation.

For other details, see the supporting material.

```
### Supporting material:
DCS_2022_Q27_Configurable_Optional_Features_File_Management_data{.ts
  v,.xlsx}
DCS_2022_Q27_Configurable_Optional_Features_File_Management_chart.pn
  g
```

Question 28. Configurable/Optional Features: Metadata Management

In Question 28, respondents were asked to indicate to what extent their Dataverse installations use features/tools to support metadata management.

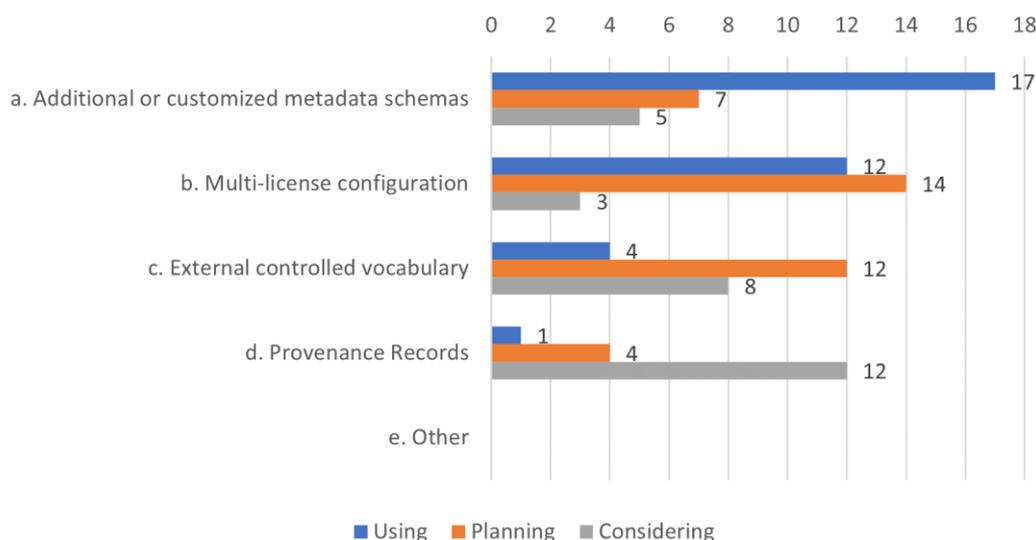Figure 23 shows the number of submitted responses for three selected response options.

Figure 23: Results from Question 28. Configurable/Optional Features: Metadata Management. The figure shows the number of submitted responses for three selected response options.

Half of the 34 Dataverse installations participating in the survey use additional or customized metadata schemas and 12 installations are planning to use or considering using this feature. Support for multiple standard licenses was introduced in March 2022 with version 5.10 of the Dataverse software, thus only a couple of months before this survey was carried out. Despite this recent introduction, this feature is already used by 12 installations and planned to or considered being used by another 17 installations. Similarly, experimental support for external controlled vocabularies was introduced with version 5.7 in October 2021. Somewhat more than half a year later, when this survey was conducted, 4 installations were using, and 20 installations were planning to use or considering using this feature. Only one installation is using the possibility to add provenance records to datasets, whereas this feature is planned to be or considered being used by 16 installations.

For other details, see the supporting material.

```
### Supporting material:
DCS_2022_Q28_Configurable_Optional_Features_Metadata_Management_data
   {.tsv,.xlsx}
DCS_2022_Q28_Configurable_Optional_Features_Metadata_Management_char
   t.png
```

Question 29. Configurable/Optional Features: Other

In Question 29, respondents were asked to indicate to what extent their Dataverse installations use other types of features/tools than the ones covered in Questions 21 to 28 above.

Figure 24 shows the number of submitted responses for three selected response options.
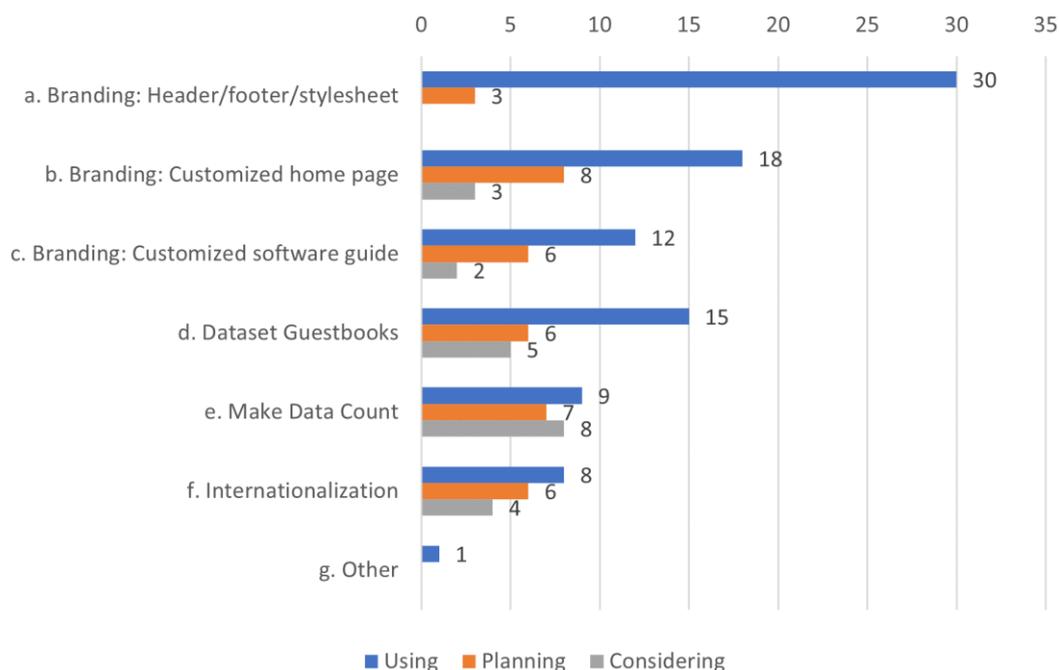
Figure 24: Results from Question 29. Configurable/Optional Features: Other. The figure shows the number of submitted responses for three selected response options.

Of the 34 Dataverse installations participating in the survey, 30 use the possibility to customize the header/footer/stylesheet to brand their installations. Customization of the homepage has been used by 18 installations and is planned to be or considered being used by another 11 installations. The possibility to customize the Dataverse software guide is used by 12 installations and planned to be or considered being used by 8 installations. 15 installations have activated the dataset guestbook feature, whereas 11 installations are planning to or considering using this feature. Make Data Count support is deployed by 9 installations and planned to be or considered being used by another 15 installations. Internationalization support is used by 8 installations and 12 installations are planning to use or considering using this feature.

One installation mentioned in the *Other* field that they are using an external data request system integrated through a fork of the Dataverse software code.

For other details, see the supporting material.

```
### Supporting material:
DCS_2022_Q29_Configurable_Optional_Features_Other_data{.tsv,.xlsx}
DCS_2022_Q29_Configurable_Optional_Features_Other_chart.png
```

Community Engagement

Questions 30 and 31 examine to what extent and in which way(s) the installations engage in the Dataverse community. Both questions are closed-ended single choice questions with the possibility to elaborate in an open-ended / free-text field.

Question 30. GDCC Membership: Is your Dataverse installation / your organization a member of the Global Dataverse Community Consortium?

Question 30 asks respondents to indicate the GDCC membership status of their Dataverse installation. As mentioned earlier, the vision of the Global Dataverse Community Consortium (GDCC) is to provide international organization to existing Dataverse community efforts to support Dataverse repositories around the world.

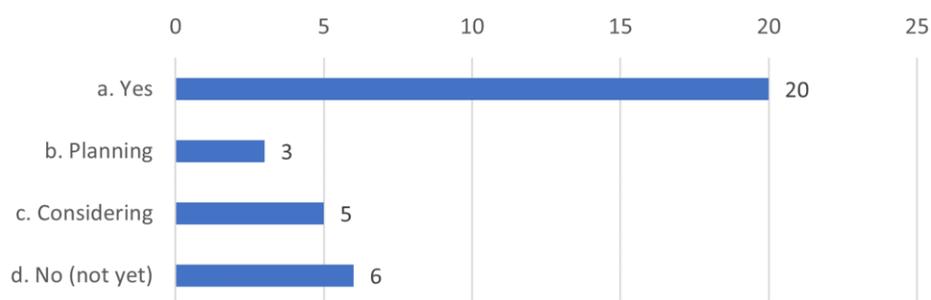Figure 25 shows the number of submitted responses for each selected response option.

Figure 25: Results from Question 30. GDCC Membership: Is your Dataverse installation / your organization a member of the Global Dataverse Community Consortium? The figure shows the number of submitted responses for each selected response option.

A large majority of the Dataverse installations participating in the survey are already a member or are planning to become or considering becoming a member of the GDCC.

```
### Supporting material:
DCS_2022_Q30_GDCC_Membership_data{.tsv,.xlsx}
DCS_2022_Q30_GDCC_Membership_chart.png
```

Question 31. Contribution: Has your organization contributed to the development and maintenance of Dataverse software code or other community efforts?

In this question, respondents were asked to indicate whether their Dataverse installation has contributed to the development and maintenance of the Dataverse software, either with software code, or by contributing to other community efforts.

Figure 26 shows the number of submitted responses for the selected response options.

Of the 34 Dataverse installations participating in the survey, 8 have contributed with software code, 10 have contributed to other community efforts, and 7 have contributed with both software code and other efforts.
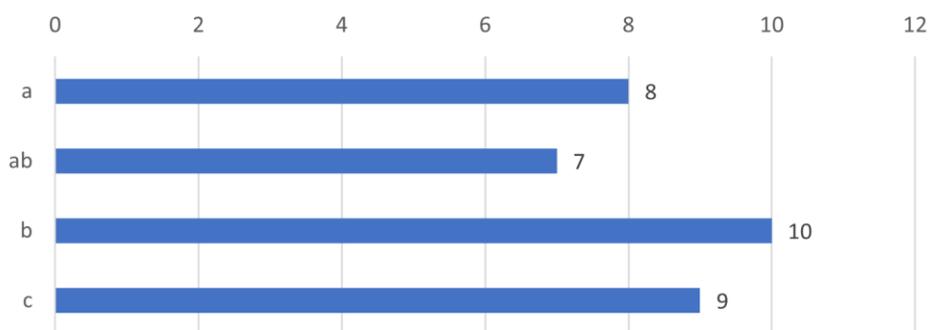
Figure 26: Results from Question 31. Contribution: Has your organization contributed to the development and maintenance of Dataverse software code or other community efforts? The figure shows the number of submitted responses for the selected response options. Abbreviations: a = Yes, to software code; b = Yes, to other community efforts; c = No (not yet).

```
### Supporting material:
DCS_2022_Q31_Contribution_data{.tsv,.xlsx}
DCS_2022_Q31_Contribution_chart.png
DCS_2022_Q31_Contribution_type_of_response_data{.tsv,.xlsx}
```

PART 4: ROADMAP AND PRIORITIES

Question 32. Roadmap: Does your Dataverse installation have a roadmap or strategic plan?

Question 32 is a closed-ended single choice question asking respondents to indicate whether their Dataverse installation has a roadmap or strategic plan. Figure 27 shows the number of submitted responses for each selected response option.
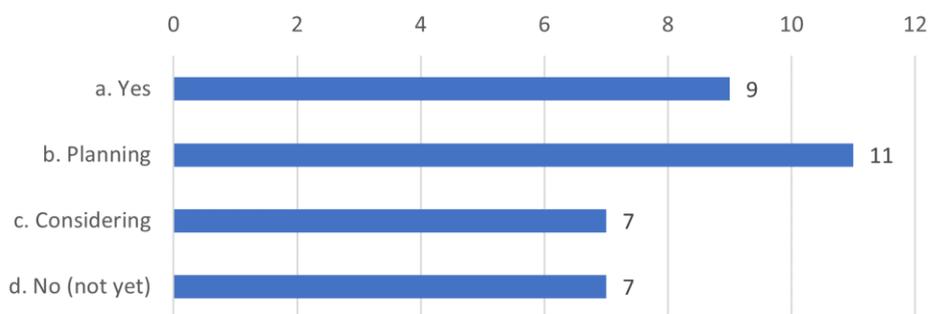

Figure 27: Results from Question 32. Roadmap: Does your Dataverse installation have a roadmap or strategic plan? The figure shows the number of submitted responses for each selected response option.

27, i.e., more than two thirds of the Dataverse installations participating in the survey either have, are planning to have or considering having a roadmap or strategic plan.

Some of the installations already having a roadmap or strategic plan also added the URL in the survey form; see column "Q33_Roadmap_URL" in the anonymized raw data file "DCS_2022_data{.tsv,.xlsx}".

```
### Supporting material:
DCS_2022_Q32_Roadmap_data{.tsv,.xlsx}
```

`DCS_2022_Q32_Roadmap_chart.png`

Support provided / to be provided by the Dataverse software

The aim of this section of the survey is to get an overview of how important different areas of support provided by the Dataverse software are considered by Dataverse installations. Questions 34 to 38 are all optional closed-ended multiple choice questions, where respondents for each support area could select between the following response options: "Not applicable", "Not important", "Somewhat important", "Neither nor", "Important", "Very important". In addition to the listed support areas, respondents were given the possibility to add their own alternative(s) and/or elaborate on other matters in an open-ended / free-text field.
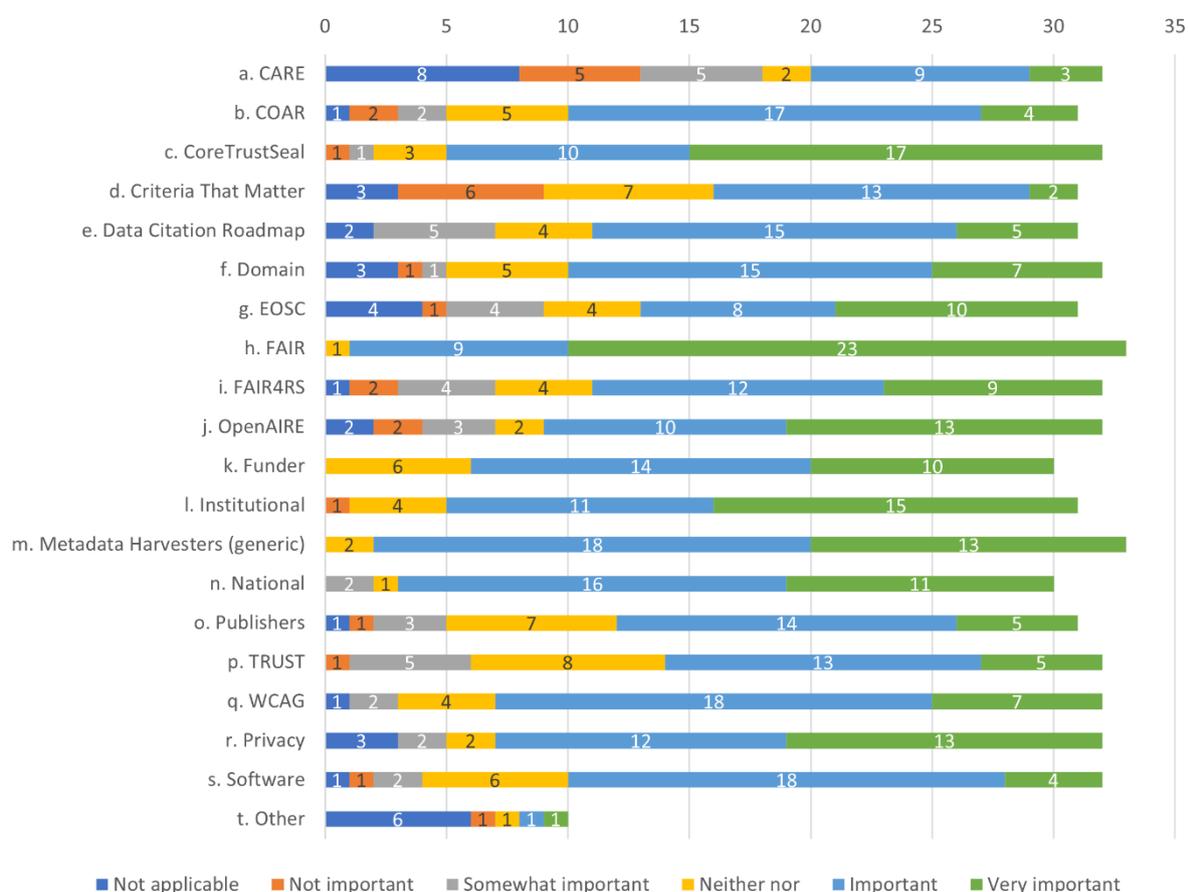
Question 34. Alignment and Compliance: How important is it for your Dataverse installation that the Dataverse software supports / should support the alignment or compliance with the following policies, requirements, guidelines, and recommendations?

In Question 34, respondents were asked about the importance of the Dataverse software supporting the alignment or compliance with the following 20 (types of) policies, requirements, guidelines, and recommendations:

a. CARE: CARE Principles for Indigenous Data Governance (https://www.gida-global.org/care)
b. COAR: COAR Community Framework for Good Practices in Repositories (https://www.coar-repositories.org/coar-community-framework-for-good-practices-in-repositories/)
c. CoreTrustSeal: Core Trustworthy Data Repositories Requirements (https://www.coretrustseal.org/wp-content/uploads/2019/02/CoreTrustSeal-Draft_Requirements_2020-2022_v00_01.pdf)
d. Criteria That Matter: Data Repository Selection: Criteria That Matter (https://osf.io/m2bce/)
e. Data Citation Roadmap: Data citation roadmap for scholarly data repositories (https://doi.org/10.1038/s41597-019-0031-8)
f. Domain: Domain/community-specific policies, requirements, guidelines, and recommendations
g. EOSC: EOSC Provider Portal Inclusion Criteria (https://eosc-portal.eu/providers-documentation/eosc-provider-portal-inclusion-criteria); EOSC Interoperability Framework (https://doi.org/10.2777/620649)
h. FAIR: FAIR data principles (https://www.go-fair.org/fair-principles/)
i. FAIR4RS: FAIR Principles for Research Software (https://doi.org/10.15497/RDA00068)
j. OpenAIRE: OpenAIRE Guidelines for Data Archives (https://guidelines.openaire.eu/en/latest/data/index.html)
k. Funder: Funder policies, requirements, guidelines, and recommendations
l. Institutional: Institutional policies, requirements, guidelines, and recommendations

m. Metadata Harvesters (generic): e.g., B2FIND, BASE Bielefeld, DataCite, Google Dataset Search
n. National: National policies, requirements, guidelines, and recommendations
o. Publishers: Publisher policies, requirements, guidelines, and recommendations
p. TRUST: TRUST Principles for digital repositories (https://doi.org/10.1038/s41597-020-0486-7)
q. WCAG: Web Content Accessibility Guidelines (https://www.w3.org/TR/WCAG21/)
r. Privacy: E.g., GDPR.
s. Software: Software Citation Principles (https://peerj.com/articles/cs-86)
t. Other: Specify below

Figure 28 shows the number of submitted responses for each selected response option.

Figure 28: Results from Question 34. Alignment and Compliance: How important is it for your Dataverse installation that the Dataverse software supports / should support the alignment or compliance with the following policies, requirements, guidelines, and recommendations? The figure shows the number of submitted responses for each selected response option.

In the next figure, Figure 29, response options other than "Important" and "Very important" are excluded and the response options "Important" and "Very important" are merged. The support options are sorted in descending order with the option considered as important/very important by most Dataverse installations being placed at the top the chart.
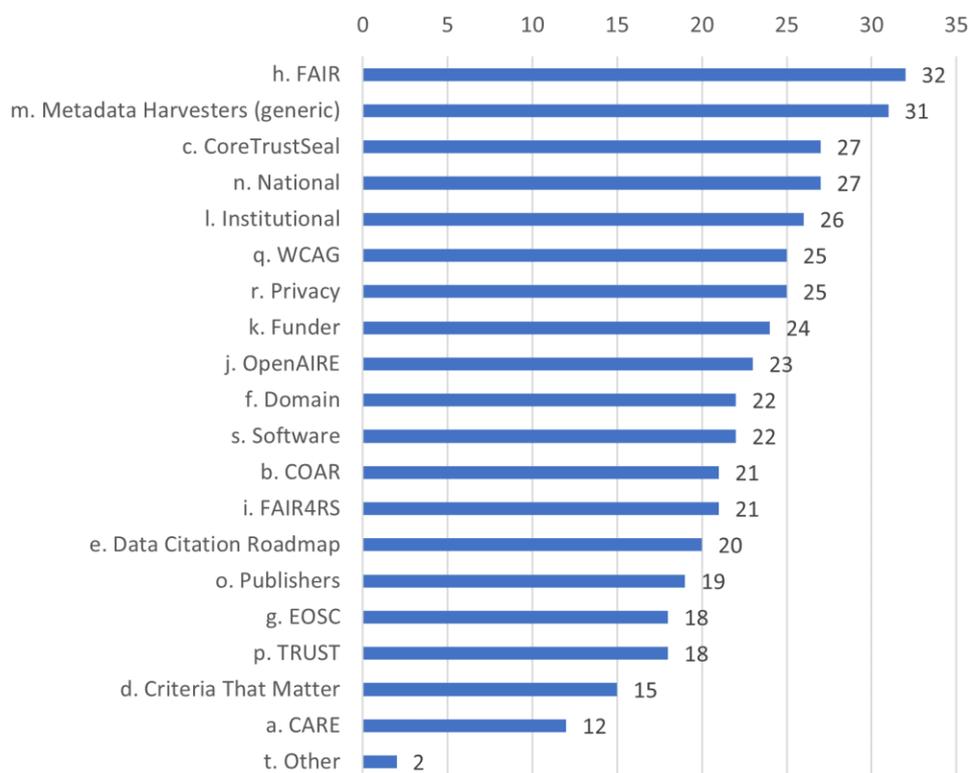
Figure 29: Rearrangement of responses to Question 34. Alignment and Compliance : How important is it for your Dataverse installation that the Dataverse software supports / should support the alignment or compliance with the following policies, requirements, guidelines, and recommendations? The figure shows the sum of the number of submitted responses for the response options "Important" and "Very important". The support options are sorted in descending order with the option considered as important/ very important by most Dataverse installations being placed at the top the chart.

If we disregard "t. Other", we can conclude that at least half of the Dataverse installations responding about a particular option (i.e., a policy, a (set of) requirement(s), a guideline(s), or a recommendation(s) consider that option to be important or very important to be supported by the Dataverse software, except for the two options "d. Criteria That Matter" and "a. CARE", for which the share of "important" / "very important" responses is lower than half of the submitted responses. The five areas for which alignment/compliance support is considered to be important or very important by most respondents are requirements or recommendations regarding "h. FAIR", "m. Metadata Harvesters (generic)", "c. CoreTrustSeal", "n. National", and "l. Institutional".

For other details, the reader is referred to the supporting material.

```
### Supporting material:
DCS_2022_Q34_Alignment_and_Compliance_data{.tsv,.xlsx}
DCS_2022_Q34_Alignment_and_Compliance_chart.png
DCS_2022_Q34_Alignment_and_Compliance_Important_data{.tsv,.xlsx}
DCS_2022_Q34_Alignment_and_Compliance_Important_chart.png
```

Question 35. Repository Features - Importance?: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas

Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature.

The next two questions are closed-ended multiple choice questions about a selection of repository features, but from two different angles.

In Question 35, respondents were asked to indicate how important it is for their Dataverse production installation that the Dataverse software supports / should support different repository features.

In the subsequent question, Question 36, the respondents were asked to indicate which architecture choice(s) they prefer for each of the features. The respondents could choose one of more of three major architecture choices, which were briefly illustrated in a linked Google document, repeated below in Table 1:

Table 1: Three Architecture Choices for Repository Features

Architecture Choice	Current Examples	Pros	Cons
a. Out of the box	• Authentication • Checksums • Embargo • Provenance • Versioning	• Things work more or less out of the box. • Has a few moving parts (e.g., one process, one app server, one database). As a result, it is easier to design, deploy, and test (system test, e2e test) the application. • Easy to manage transactions and data. sharing between features • Low operational complexity.	• Only configurable, cannot remove it. • Hard to adapt (fork or live with it). • Difficult to parallelize work among multiple teams. So, development scaling is challenging. • Granular scaling (i.e., scaling part of the application) is not possible. • Polyglot programming or polyglot databases are challenging.
b. Extension	• External controlled vocabularies • Previewers • Localization/	• Better development scaling as teams can work parallely on different features in a more autonomous	• Relies on community maintenance. • Difficult to coordinate.

Architecture Choice	Current Examples	Pros	Cons
	• Internationali-zation	way with little external dependency, thus good support for crowdsourcing. • Can be pluggable.	• Harder to sustain/sync the different parts. • Breaking changes. • Less seamless UI.
c. Loosely coupled integration (via API)	• Archivematica • Data Curation Tool • Open Journal System • Whole Tale	• Greatest degree of flexibility and freedom.	• Relies on external maintenance. • Same as choice b), but stronger.

In addition to these three architecture choices, there were two more response options: "No preferences" and "Other; specify below".

Question 35 (Importance) and Question 36 (Architecture Choices) cover the same 54 repository features, which group into the following 14 topics:

- Core functionality
- RDM planning
- Deposit and Curation
- Large Data Support
- Metadata Support
- File Management and Reuse Support
- Data Privacy Support
- Documentation and Analysis Support
- Preservation Support
- Discovery and Reuse Support
- Accessibility Support
- Internationalization Support
- DevOps Support
- Scalability Support

Below, the results from Question 35 and Question 36 will be discussed together, one feature at a time, grouped into sections according to the 14 topics. For other details, the reader is referred to the supporting material.

Core Functionality
In the survey form, the following three repository features are included in the topic Core Functionality:

a. Authentication (e.g., OAuth, OpenID Connect, Shibboleth) and Authorization (e.g., groups, roles)
b. PIDs (e.g., DOI, Handle, ORCID, RAiD, ROR)
c. (Almost) UI-equivalent API

Figure 30 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for the repository features included in the topic Core Functionality.
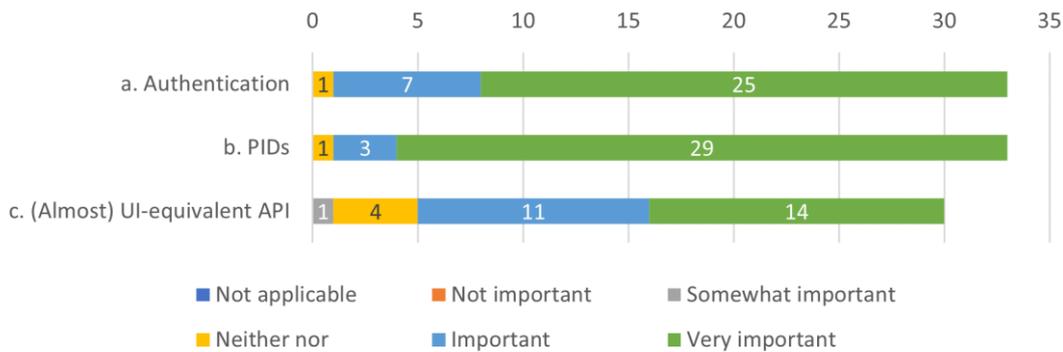

Figure 30: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository features within the the topic Core Functionality.

Support for all three features is considered being important or very important by a large majority of the Dataverse installations participating in the survey.

Figure 31 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository features included in the topic Core Functionality.
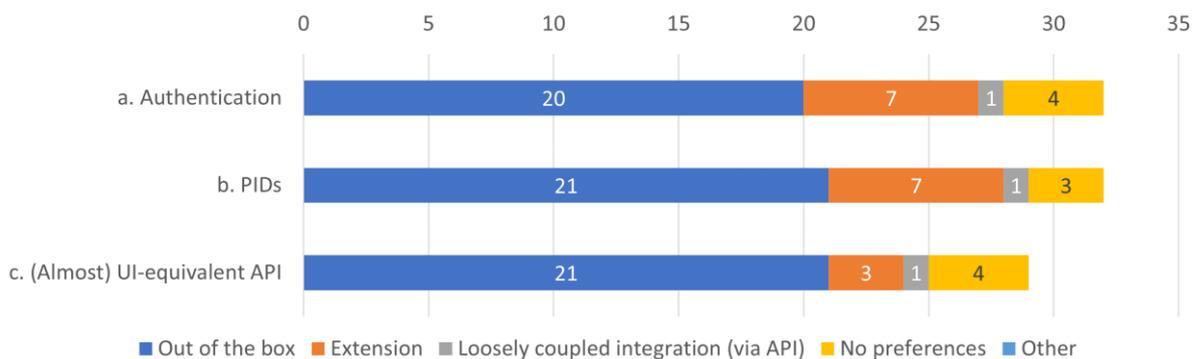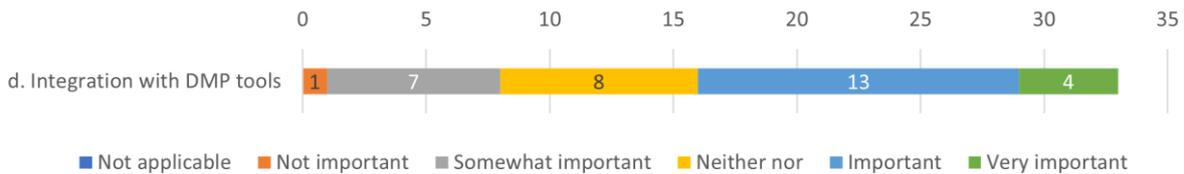

Figure 31: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic Core Functionality.

A majority of the Dataverse installations participating in the survey prefer these three Core Functionality features to be implemented out of the box.

RDM Planning Support

The topic RDM Planning Support only includes one repository feature: "d. Integration with DMP tools (e.g., Argos, Data Stewardship Wizard, DMPonline, DMPTool, easyDMP)".

Figure 32 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for this repository feature.

Figure 32: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository feature within the the topic RDM Planning Support.

About half of the respondents answering this part of Question 35 consider support for integration with DMP tools as important or very important.

Figure 33 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository feature "Integration with DMP tools".
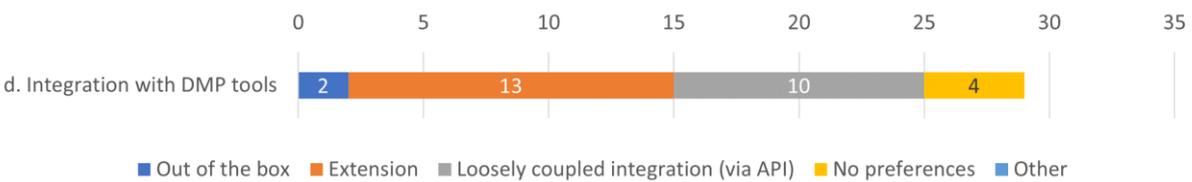

Figure 33: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository feature within the topic RDM Planning Support.

A majority of the responding Dataverse installations prefer support for integration with DMP tools to be implemented as extension or as loosely coupled integration (via API).

Deposit and Curation Support

The topic Deposit and Curation Support covers the following two repository features in the survey:

e. Deposit and Curation - Depositing facilitation/automation
f. Deposit and Curation - Curation management

Figure 34 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for the repository features within the topic Deposit and Curation Support.
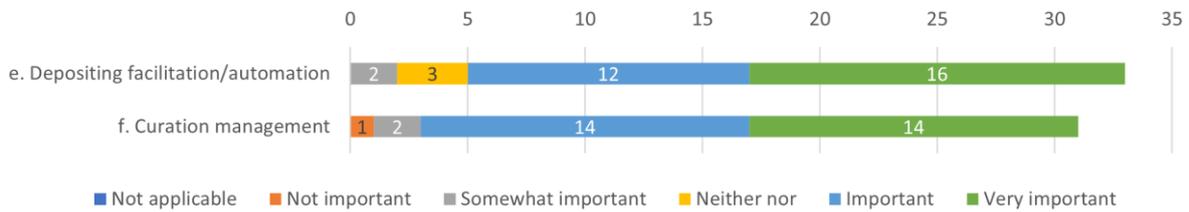

Figure 34: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository features within the topic Deposit and Curation Support.

Support for both features within the topic Deposit and Curation Support is considered being important or very important by a large majority of the Dataverse installations participating in the survey.

Figure 35 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository features included in the topic Deposit and Curation Support.
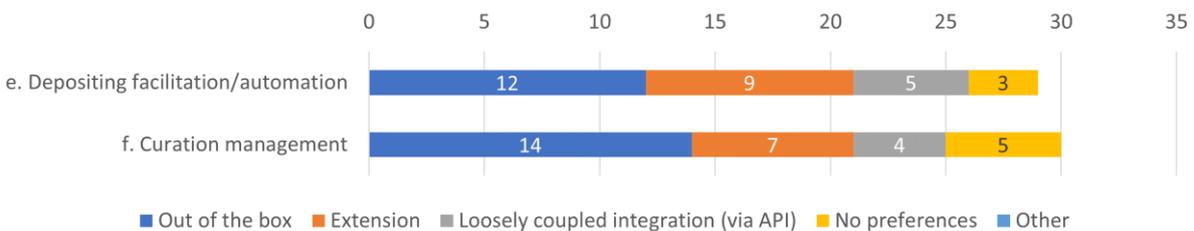

Figure 35: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic Deposit and Curation Support.

Among the Dataverse installations participating in the survey and answering to this part of Question 36, "out of the box" is chosen most often as the preferred architecture choice for these two Deposit and Curation support features, but also implementation as an extension is preferred by a considerable share of installations responding to this part of the question.

Large Data Support

There are two aspects of Large Data Support that are covered in the survey: size per file, and number of files per dataset. The respondents were asked to indicate their opinion

on importance and architecture choice(s) for different ranges of file size and number of files.

For file size, the following ranges are used in the survey:

g. Up to 4 GB
h. Up to 10 GB
i. Up to 100 GB
j. Up to 1 TB
k. Up to 10 TB
l. More than 10 TB

Figure 36 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for the different ranges of file size within the topic Large Data Support.
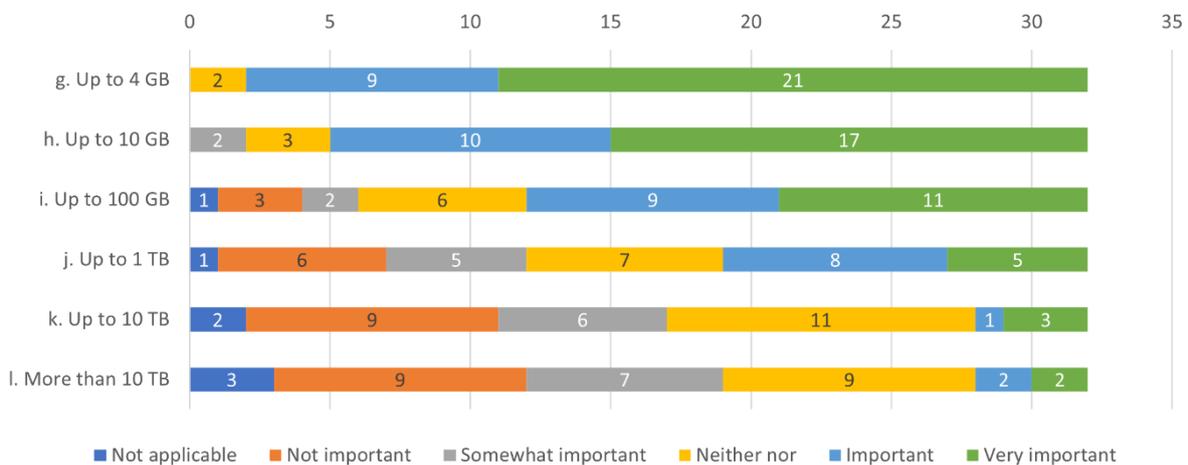

Figure 36: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository features within the topic Large Data Support - File Size.

We can observe that support for large data is considered less important as the file size increases. Support for file sizes up to 100 GB is important or very important for more than half of the Dataverse installations participating in the survey. Support for file sizes up to 1 TB is considered as somewhat important, important, or very important by more than half of the installations, whereas support for larger files is considered not important, or not applicable.

Figure 37 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the different ranges of file size within the topic Large Data Support.
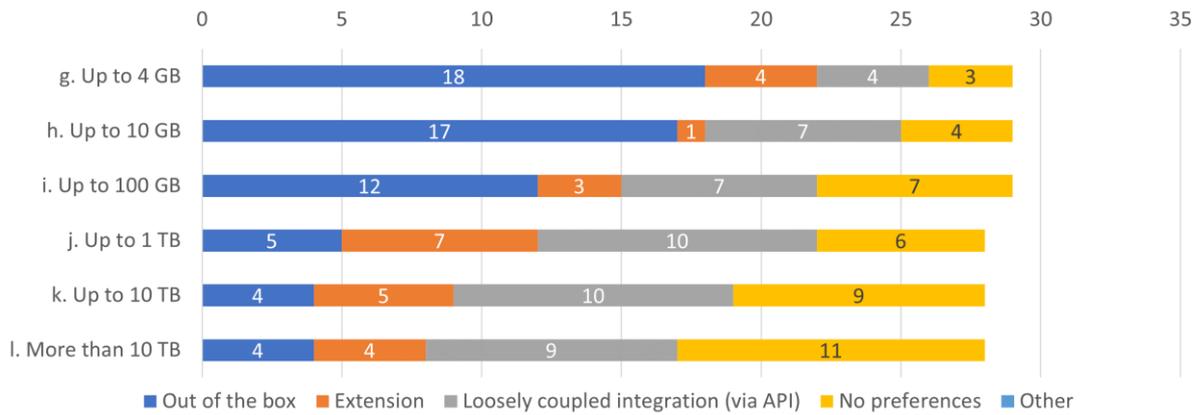
Figure 37: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic Large Data Support - File Size.

For supporting file sizes up to 100 GB, "out of the box" is the most preferred architecture choice, whereas for larger file sizes, both "extension" and "loosely coupled integration (via API)" are equally or more preferred choices.

For number of files per dataset, the following ranges are used in the survey:

m. Up to 100-200
n. Up to 1000
o. Up to 10000
p. Up to 1 million
q. More than 1 million

Figure 38 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for the different ranges of numbers of files per dataset within the topic Large Data Support.

In a similar way as with file size, we can observe that Large Data Support is considered less important as the number of files per dataset increases. Support for up to 1000 files per dataset is important or very important for a quite large majority of the Dataverse installations participating in the survey. Support for up to 10000 files per dataset is considered as somewhat important, important, or very important by about half of the participating installations, whereas support for more files per dataset is considered as not important, or as not applicable.
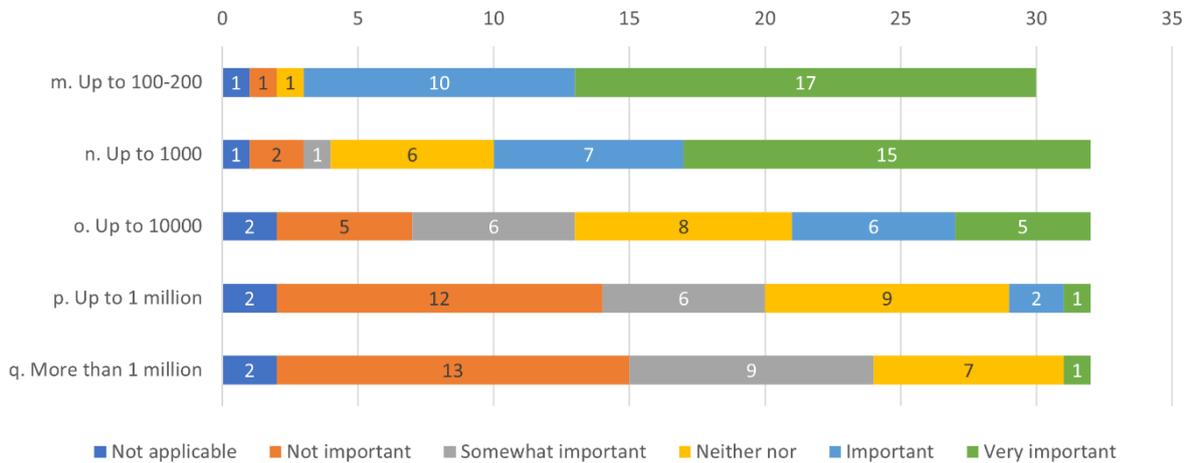
Figure 38: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository features within the topic Large Data Support - Number of Files.

Figure 39 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the different ranges of number of files per dataset within the topic Large Data Support.
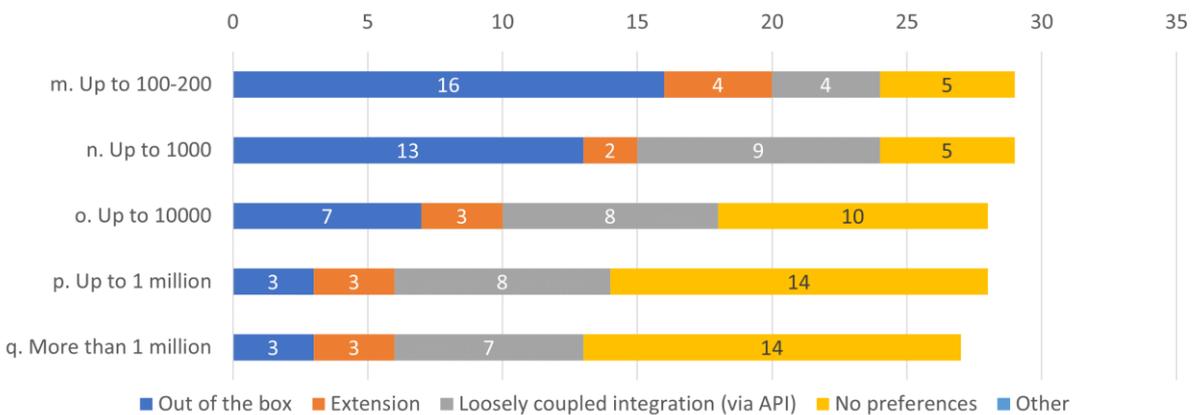

Figure 39: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic Large Data Support - Number of Files.

For supporting up to 1000 files per dataset, "out of the box" is the most preferred architecture choice, whereas for larger numbers of files per dataset, the most often selected response is "no preferences".

Metadata Support

The next topic investigated in Question 35 and Question 36 is Metadata Support, covered by the following five repository features:

r. Standardized generic metadata capturing, validation, and harvesting
s. Standardized domain-specific metadata capturing, validation, and harvesting
t. Customized metadata capturing, validation, and harvesting
u. Metadata capturing/management early in the RDM lifecycle (e.g., CEDAR)
v. External controlled vocabularies (e.g., SKOSMOS)

Figure 40 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for the repository features within the topic Metadata Support.
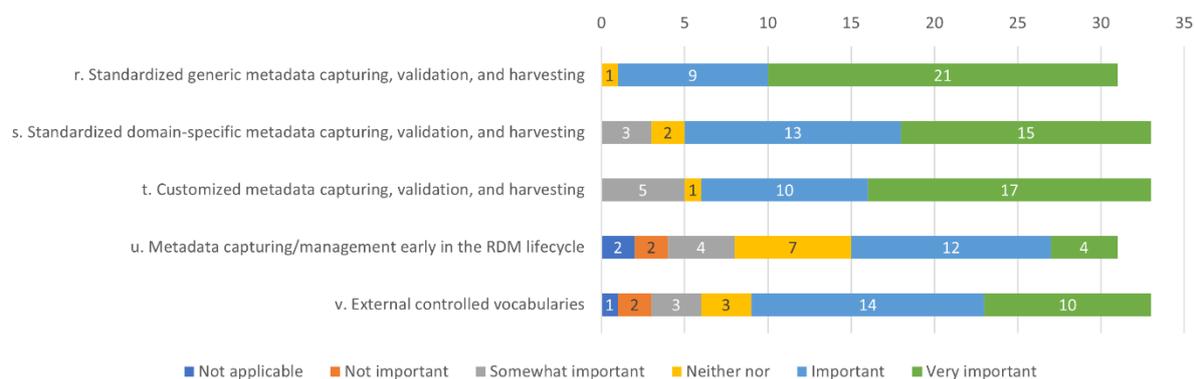

Figure 40: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository features within the topic Metadata Support.

All the included Metadata Support features are considered being important or very important by a quite large majority of the Dataverse installations participating in the survey, except for the feature "u. Metadata capturing/management early in the RDM lifecycle", which has a somewhat lower score, but is still considered as important/very important by virtually half of the installations.

Figure 41 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository features within the topic Metadata Support.

There are two repository features for which "out of the box" is the most preferred architecture choice among the Dataverse installations responding to this part of Question 36: "r. Standardized generic metadata capturing, validation, and harvesting" and "v. External controlled vocabularies". For the remaining features within the topic Metadata Support, extension is the most preferred architecture choice.
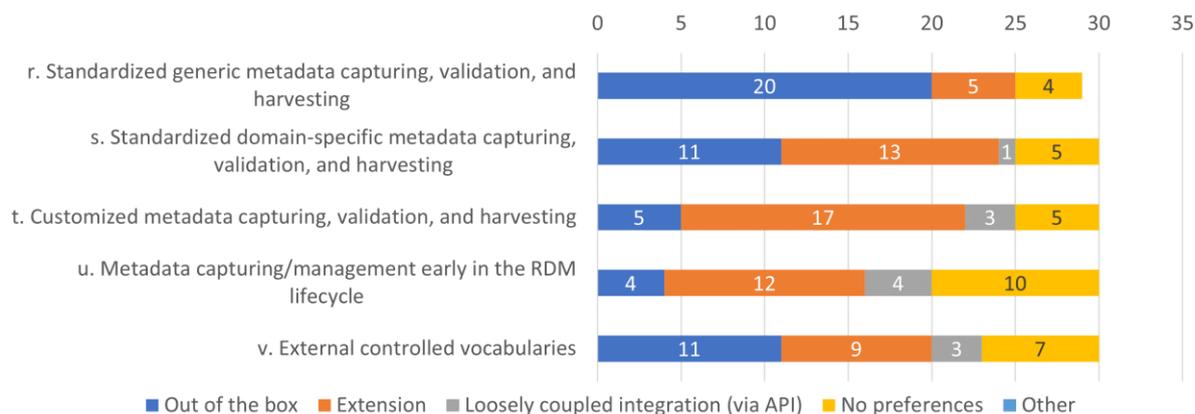
Figure 41: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic Metadata Support.

File Management and Reuse Support

The following four features are about File Management and Reuse Support:

w. Integration with (active) file storage and policy management (e.g., CKAN, GitHub, GitLab, Globus, iRODS)

x. Specialized data server technology and APIs (e.g., GeoServer, THREDDS, OPeNDAP)

y. Trusted remote storage of archived files (e.g., TRSA)

z. Virus scanning

Figure 42 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for the repository features within the topic File Management and Reuse Support.

Two of these features, "w. Integration with (active) file storage and policy management" and "z. Virus scanning" are considered as important or very important by more than half of the Dataverse installations participating in the survey. For the other two features, the scores for "important" and "very important" are somewhat lower, but both features are considered as somewhat important, important, or very important by more than half of the installations.
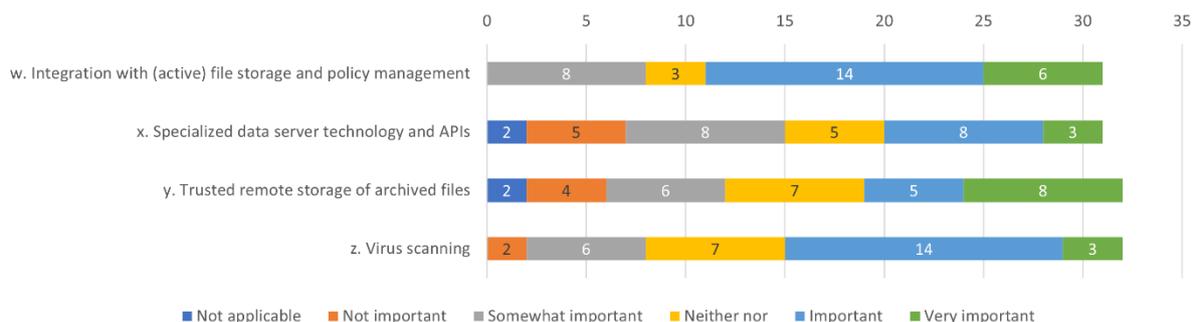
Figure 42: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository features within the topic File Management and Reuse Support.

Figure 43 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository features within the topic File Management and Reuse Support.
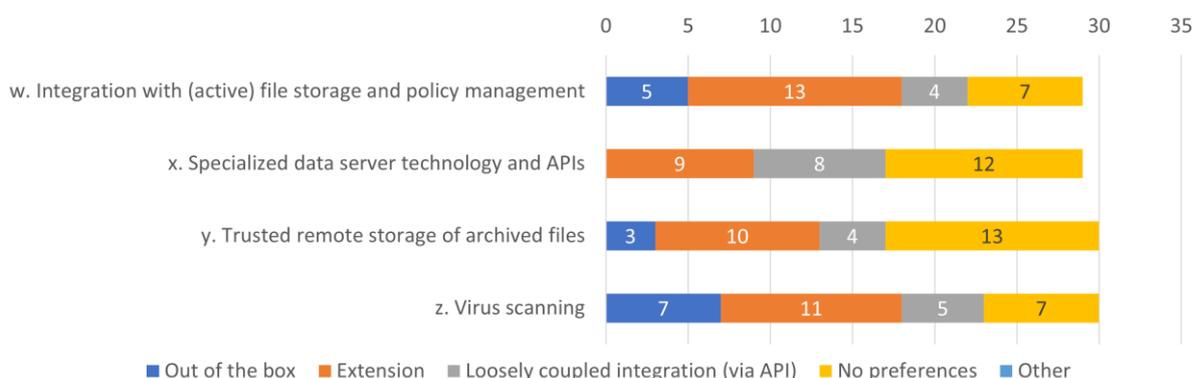

Figure 43: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic File Management and Reuse Support.

The different architecture choices are rather equally preferred for these four repository features. Extension is the most preferred choice for "w. Integration with (active) file storage and policy management" and "z. Virus scanning", whereas "No preferences" is the most selected response option for the other two features within the topic File Management and Reuse Support.

Data Privacy Support

The topic Data Privacy Support is covered by the following two repository features:

aa. Data Privacy management (e.g., DataTags, OpenDP)

ab. Trusted remote storage of archived files (e.g., TRSA)

Figure 44 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for these two repository features.
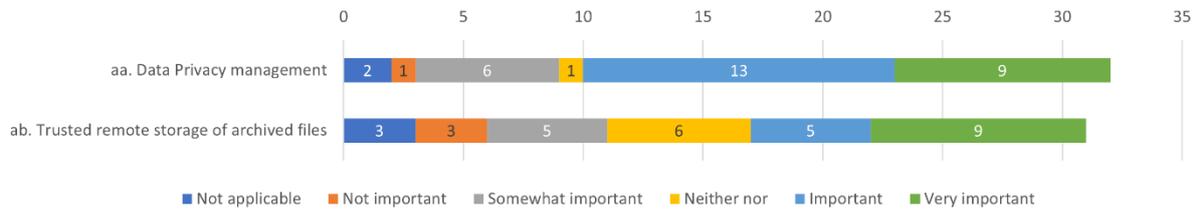

Figure 44: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository features within the topic Data Privacy Support.

Support for Data Privacy management is considered as important or very important by a quite large majority of the Dataverse installations participating in the survey. The corresponding scores for support for trusted remote storage is somewhat lower. The quite high number of installations opting for the "Neither nor" option for this feature may indicate that the use cases of this feature are not entirely clear to these respondents.

Figure 45 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository features within the topic Data Privacy Support.
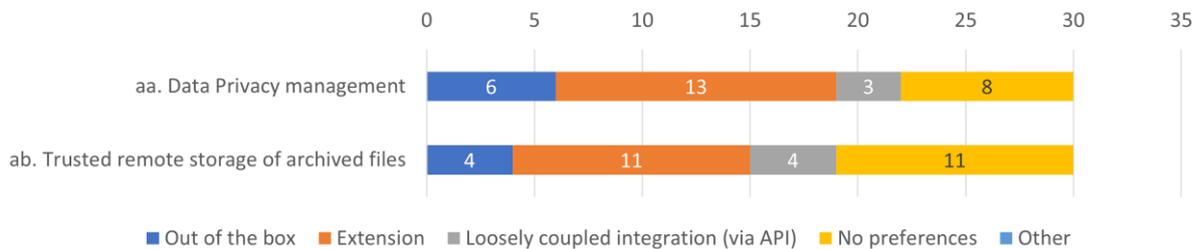

Figure 45: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic Data Privacy Support.

Among the Dataverse installations responding to this part of Question 36, extension is the most preferred architecture choice for supporting Data Privacy management, whereas for support for trusted remote storage of archived files, "Extension" and "No preferences" are selected as preferred architecture choice by an equal number of installations.

Documentation and Analysis Support

The following four features cover the topic Documentation and Analysis Support in the survey:

ac. Data analysis and computation (e.g., Apache Superset, Binder, R, Renku, Whole Tale), including Virtual Research Environments (VREs)

ad. Electronic Lab Notebooks (e.g., RSpace)

ae. Computational notebooks (e.g., Jupyter)

af. Other project collaboration tools (e.g., OSF)

Figure 46 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for these four repository features.
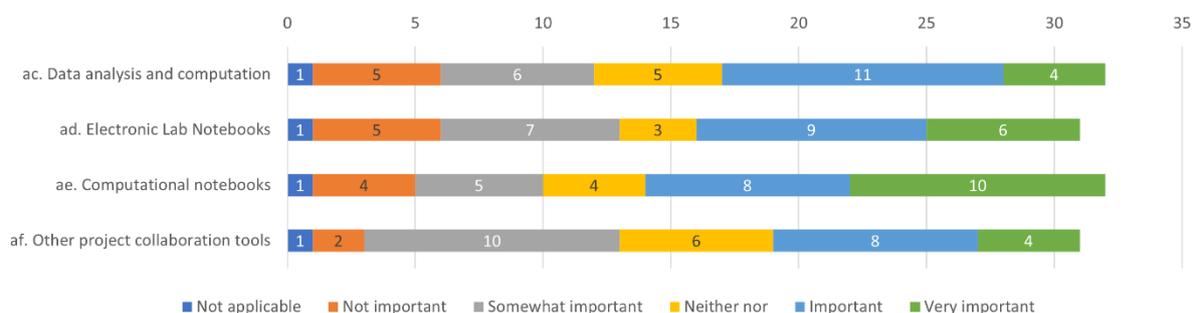

Figure 46: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository features within the topic Documentation and Analysis Support.

Support for computational notebooks is rated to be important or very important by slightly more than half of the Dataverse installations participating in the survey, whereas the corresponding scores for support for the other features are somewhat lower. If we include the response option "Somewhat important", we can conclude that a large majority of the Dataverse installations consider all four features within the topic Documentation and Analysis Support as either somewhat important, important, or very important.

Figure 47 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository features within the topic Documentation and Analysis Support.
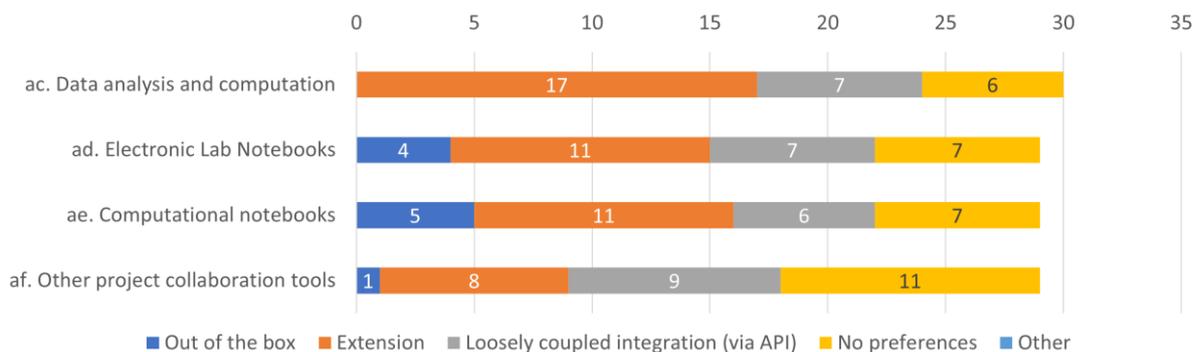
Figure 47: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic Documentation and Analysis Support.

Except for other project collaboration tools, extension is rated as the most preferred architecture choice for supporting the repository features within the topic Documentation and Analysis Support.

Preservation Support

The topic Preservation Support is covered by the following two features:

ag. Long-term preservation tools (e.g., Archivematica, DuraCloud/Chronopolis)

ah. Data migration (e.g., BagIt)

Figure 48 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for these two repository features.
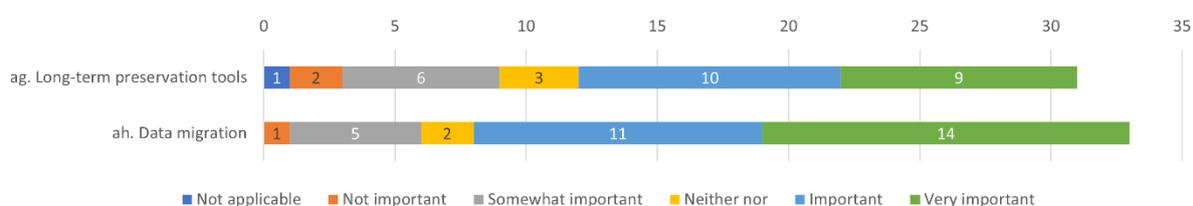

Figure 48: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository features within the topic Preservation Support.

A majority of the Dataverse installations participating in the survey consider support for both features as somewhat important, important, or very important. In all, only three installations consider these features as not important or not applicable.

Figure 49 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository features within the topic Preservation Support.
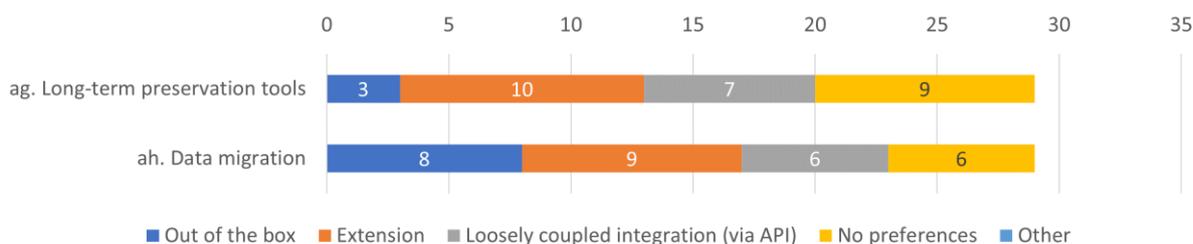
Figure 49: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic Preservation Support.

Although both features within the topic Preservation Support are most preferably seen implemented through extension, the preference is much more evenly distributed across the different architecture choices, and a considerable number of respondents did not have any preferences for any of the choices.

Discovery and Reuse Support

The following five features are about Discovery and Reuse Support:

   ai. Discoverability (e.g., Geodisy, OAI-PMH, schema.org)

   aj. Computational reproducibility (e.g., Core2)

   ak. Software encapsulation

   al. Annotation (e.g., B2NOTE, hypothes.is, Pundit)

   am. Peer reviewing and article/book publishing (e.g., Libero, OJS, OMP)

Figure 50 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for these repository features within the topic Discovery and Reuse Support.
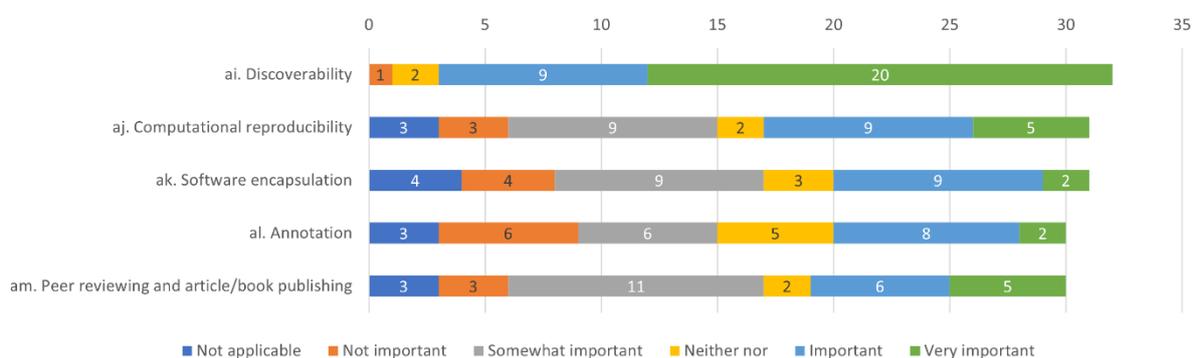

Figure 50: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository features within the topic Discovery and Reuse Support.

Support for discoverability is without doubt a repository feature which is considered as important or very important by a large majority of the Dataverse installations participating in the survey. For the other four features within the topic Discovery and Reuse Support, less than half of the installations opted for "important" or "very important", whereas quite a large share of the installations consider these features as somewhat important.

Figure 51 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository features within the topic Discovery and Reuse Support.
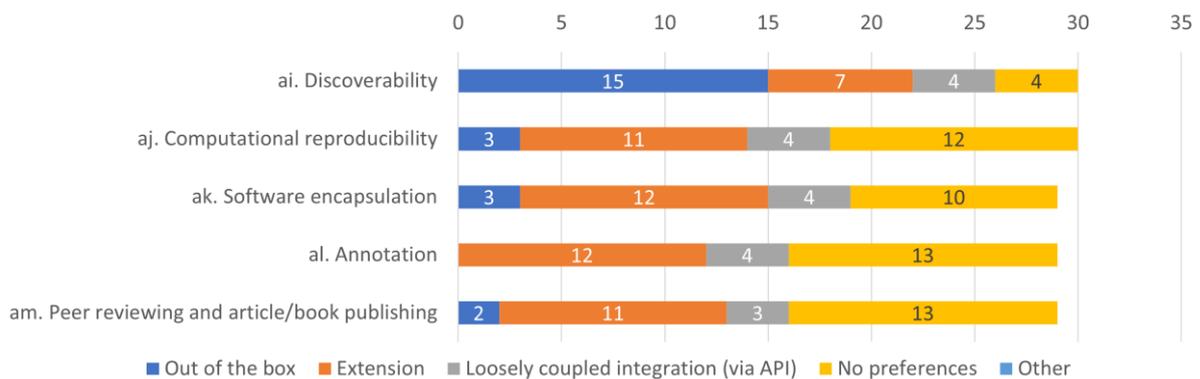

Figure 51: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic Discovery and Reuse Support.

"Out of the box" is the most preferred architecture choice for discoverability support. For the other four repository features within the topic Discovery and Reuse Support, most respondents either prefer support to be implemented through extension, or they do not have any preferences with respect to architecture choice(s).

Accessibility Support

Support for Universal/Inclusive Design (e.g., WCAG 2.0 or 2.1) is the only feature within the topic Accessibility Support.

Figure 52 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for this repository feature.

With two exceptions, all Dataverse installations responding to this part of Question 35 consider support for Universal/Inclusive Design as somewhat important, important, or very important.
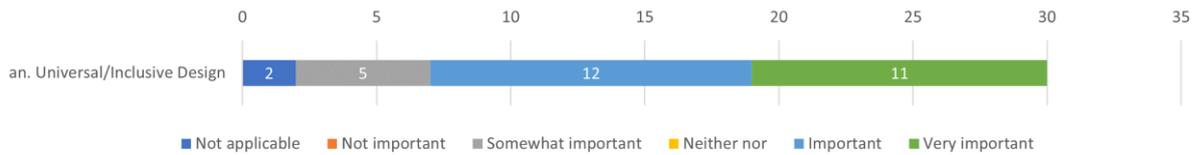
Figure 52: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository feature within the topic Accessibility Support.

Figure 53 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository features within the topic Accessibility Support.
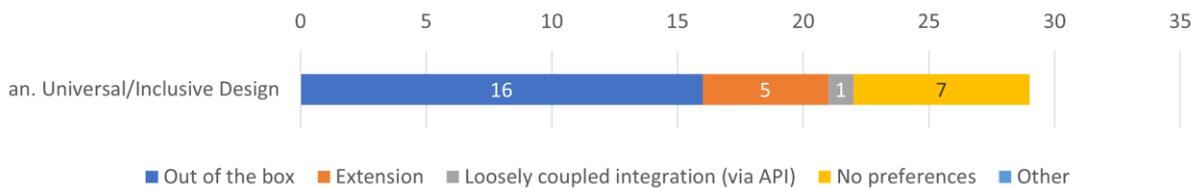

Figure 53: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic Accessibility Support.

More than half of the Dataverse installations responding to this part of Questions 36 want Universal/Inclusive Design to be supported out of the box.

Internationalization Support

Also the next topic, Internationalization Support, consists of only one repository feature, namely multiple language support.

Figure 54 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for this repository feature.
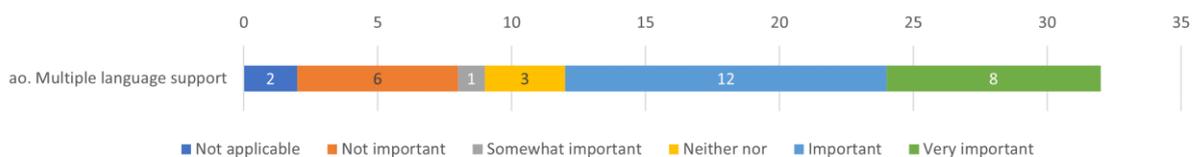

Figure 54: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository feature within the topic Internationalization Support.

Multiple language support is important or very important for a considerable majority of the Dataverse installations participating in the survey. Within the remaining installations, quite a large share (6 installations) considers this repository feature as not important.

Figure 55 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository features within the topic Internationalization Support.
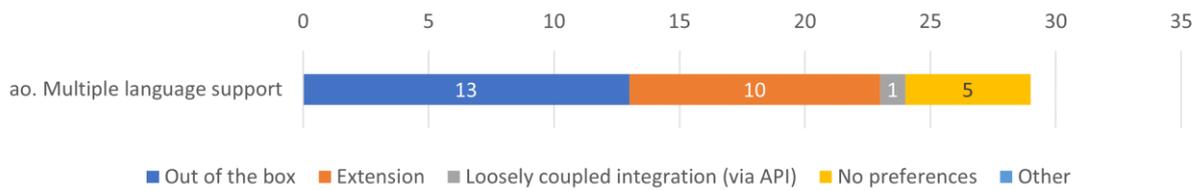

Figure 55: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository feature within the topic Internationalization Support.

"Out of the box" is the most preferred architecture choice for supporting multiple language support, closely followed by extension, which is the preferred choice for only three fewer respondents.

DevOps Support

DevOps refers to a set of practices that combines software development (Dev) and IT operations (Ops) ('DevOps', 2022). The topic DevOps Support is covered by the following features:

   ap. (Automated) scalability of deployment

   aq. Continuous integration (CI)

   ar. Compatibility with SolrCloud

   as. Containerized versions (Docker/Podman, Kubernetes)

   at. High availability (HA) configurations of Dataverse and its dependencies (Solr, PostgreSQL)

   au. Security (e.g., vulnerabilities)

Figure 56 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for these six repository features.

Security support is important or very important for a large majority of the Dataverse installations participating in the survey. The other repository features within the topic DevOps Support are considered as important or very important by a majority of installations, with the exception of support for compatibility with SolrCloud.
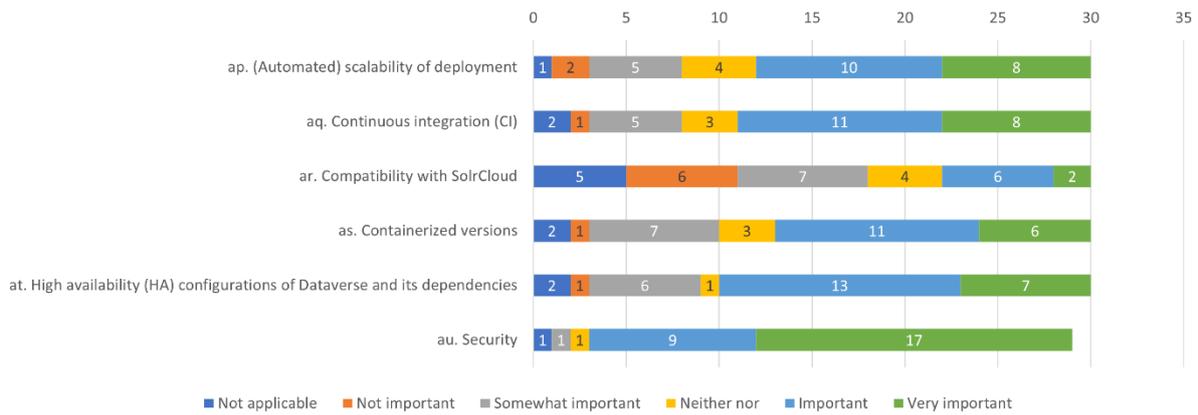
Figure 56: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository features within the topic DevOps Support.

Figure 55 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository features within the topic DevOps Support.
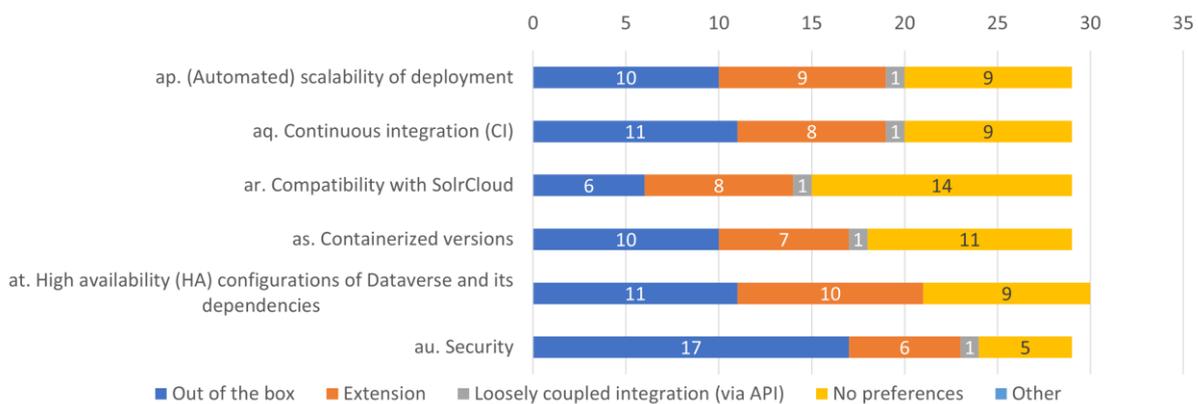

Figure 57: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic DevOps Support.

Most Dataverse installations responding to this part of Question 36 want security to be supported out of the box. Apart from security, the preferences of the installations are rather evenly distributed across the no-preferences option and the two architecture choices out of the box and extension, with the exception of compatibility with SolrCloud, for which most respondents did not have any preferences with regard to architecture choices.

Scalability Support

The following six repository features are about Scalability Support:

> av. Number of datasets
>
> aw. Files per dataset
>
> ax. File size
>
> ay. Number of metadata schemas
>
> az. Complexity of metadata schemas
>
> ba. Integrating services (e.g., through modular/microservice architecture)

For the two features "files per dataset" and "files size", there is some thematic overlap with the topic Large Data Support (see above).

Figure 58 summarizes the number of submitted responses for Question 35 (Importance) for each selected response option for these repository features within the topic Scalability Support.
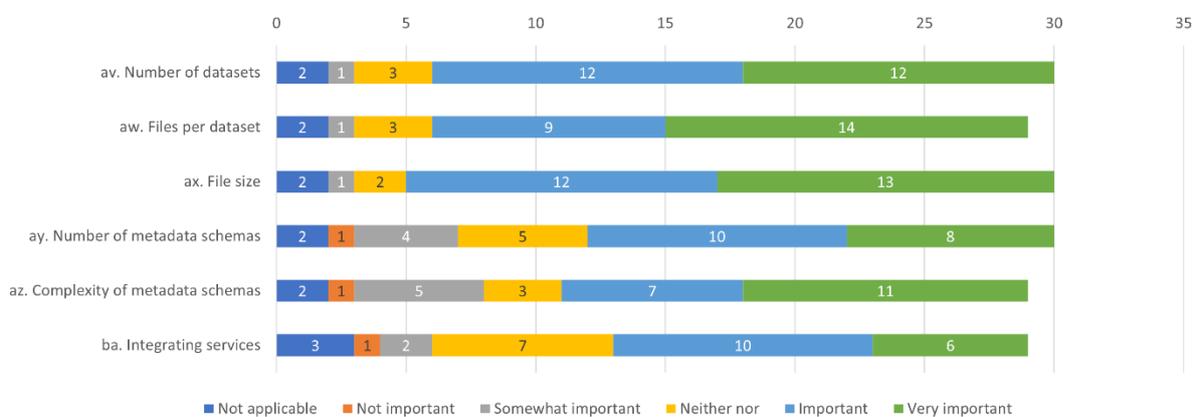

Figure 58: Results from Question 35. Repository Features - Importance: How important is it for your Dataverse production installation that the Dataverse software supports / should support features within the following areas? The figure shows the number of submitted responses for each selected response option for the repository features within the topic Scalability Support.

Scalability Support for all these six aspects is important or very important for more than half of the Dataverse installations which responded to this part of Question 35. Most important is the scalability support provided by the software in terms of file size, number of datasets within an installation, and number of files per dataset.

Figure 59 summarizes the number of submitted responses for Question 36 (Architecture Choices) for each selected response option for the repository features within the topic Scalability Support.

Except for the last feature, integrating services, out of the box is the most preferred architecture choice to support the repository features within the topic Scalability Support. However, quite a few respondents opted for extension as the preferred choice for these features, and similarly many did not have any preferences.
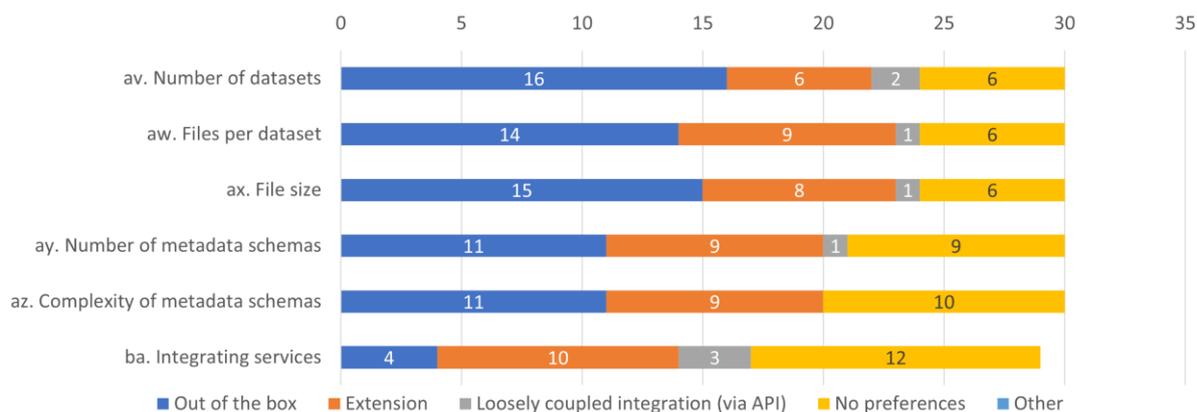
Figure 59: Results from Question 36. Repository Features - Architecture Choices: For each repository feature below, please indicate the architecture choice(s) that you prefer to be used to implement the feature. The figure shows the number of submitted responses for each selected response option for the repository features within the topic Scalability Support.

To summarize the discussion about the importance of repository features, in Figure 60 the number of submitted responses to Question 35 (Importance) for each selected response option for all features within the 14 topics are aggregated and expressed as percentages.
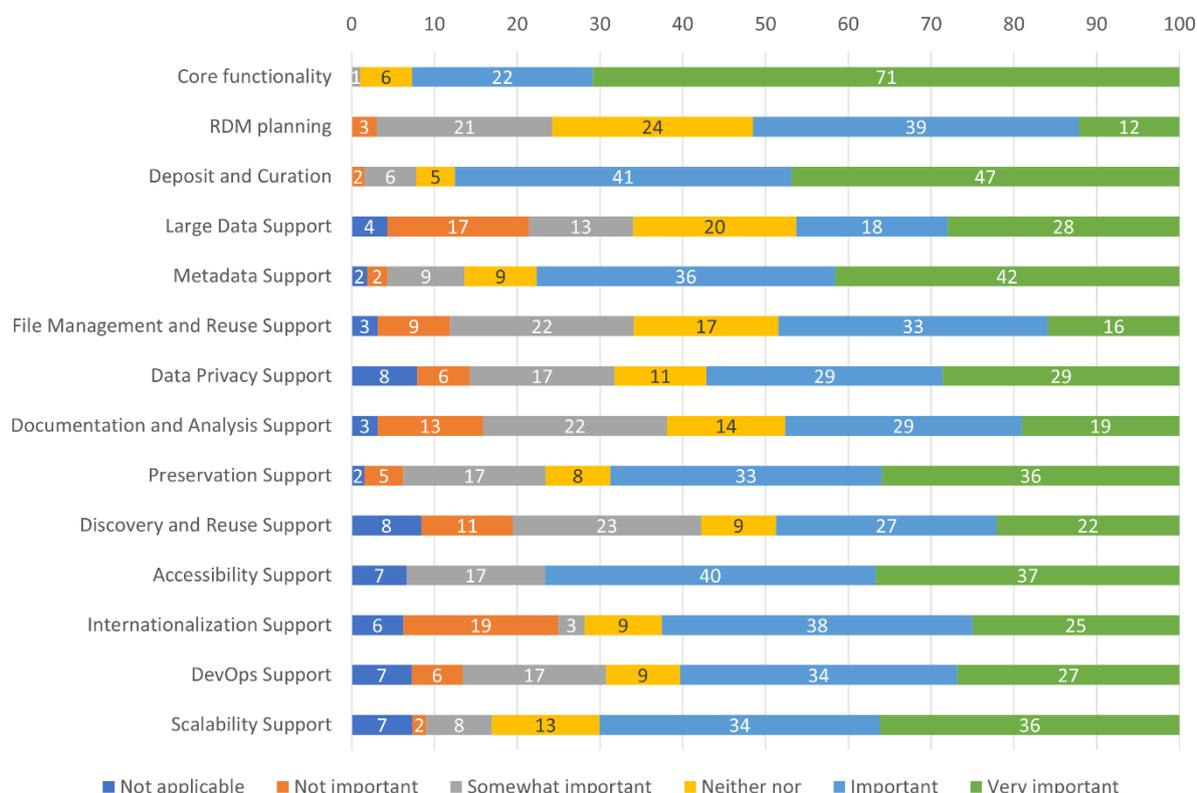

Figure 60: Importance of repository features – aggregated per topic. In the figure, the number of submitted responses to Question 35 (Importance) for each selected response option for all features within the 14 topics are aggregated and expressed as percentages.

For example, the features within the topic Core functionality were rated as very important by 71% of the installations responding to that part of Question 35.

Focusing on the top-five rated topics, we can conclude the following from Figure 60: On a aggregated level, the repository features supporting core functionality are the ones that are considered as important or very important by the highest percentage of respondents (93%), followed by the features supporting deposit and curation (88% of the respondents), the features within the topic Metadata Support (78% of the respondents), the feature supporting accessibility (77% of the respondents), and the features supporting scalability (70% of the respondents).

To summarize the discussion about the preference of architecture choices for repository features, in Figure 61 the number of submitted responses to Question 36 (Architecture Choices) for each selected response option for all features within the 14 topics are aggregated and expressed as percentages.
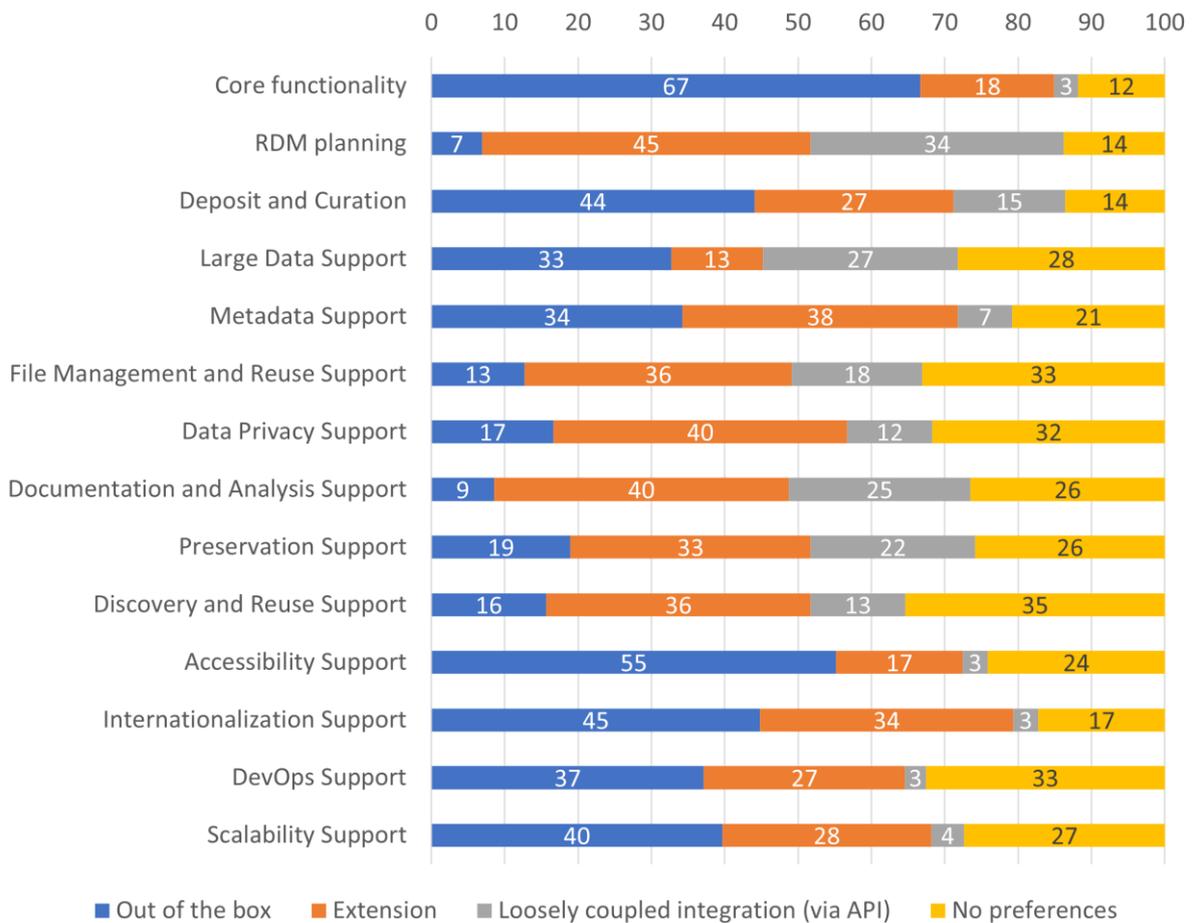

Figure 61: Preference of architecture choices for repository features – aggregated per topic. In the figure, the number of submitted responses to Question 36 (Architecture choices) for each selected response option for all features within the 14 topics are aggregated and expressed as percentages.

Of the 14 repository feature topics, core functionality is the one for which out of the box is the most preferred architecture choice (67% of the respondents), followed by Accessibility Support (55%), Internationalization Support (45%), Deposit and Curation Support (44%), Scalability Support (40%), DevOps support (37%), and Large Data Support (33%). For all other topics of repository features, extension is the most preferred architecture choice.

```
### Supporting material:
DCS_2022_Q35_Repository_Features_Importance_data{.tsv,.xlsx}
DCS_2022_Q35_Topic_01_Repository_Features_Importance_Core_Functional
   ity_chart.png
DCS_2022_Q35_Topic_02_Repository_Features_Importance_RDM_Planning_ch
   art.png
DCS_2022_Q35_Topic_03_Repository_Features_Importance_Deposit_and_Cur
   ation_chart.png
DCS_2022_Q35_Topic_04_Repository_Features_Importance_Large_Data_File
   _Size_chart.png
DCS_2022_Q35_Topic_04_Repository_Features_Importance_Large_Data_Numb
   er_of_Files_chart.png
DCS_2022_Q35_Topic_05_Repository_Features_Importance_Metadata_chart.
   png
DCS_2022_Q35_Topic_06_Repository_Features_Importance_File_Management
   _and_Reuse_chart.png
DCS_2022_Q35_Topic_07_Repository_Features_Importance_Data_Privacy_ch
   art.png
DCS_2022_Q35_Topic_08_Repository_Features_Importance_Documentation_a
   nd_Analysis_chart.png
DCS_2022_Q35_Topic_09_Repository_Features_Importance_Preservation_ch
   art.png
DCS_2022_Q35_Topic_10_Repository_Features_Importance_Discovery_and_R
   euse_chart.png
DCS_2022_Q35_Topic_11_Repository_Features_Importance_Accessibility_c
   hart.png
DCS_2022_Q35_Topic_12_Repository_Features_Importance_Internationaliz
   ation_chart.png
DCS_2022_Q35_Topic_13_Repository_Features_Importance_DevOps_chart.pn
   g
DCS_2022_Q35_Topic_14_Repository_Features_Importance_Scalability_cha
   rt.png
DCS_2022_Q35_Topic_Summary_Repository_Features_Importance_data{.tsv,
   .xlsx}
DCS_2022_Q35_Topic_Summary_Repository_Features_Importance_chart.png
DCS_2022_Q36_Repository_Features_Architecture_Choices_data{.tsv,.xls
   x}
DCS_2022_Q36_Topic_01_Repository_Features_Architecture_Choices_Core_
   Functionality_chart.png
DCS_2022_Q36_Topic_02_Repository_Features_Architecture_Choices_RDM_P
   lanning_chart.png
DCS_2022_Q36_Topic_03_Repository_Features_Architecture_Choices_Depos
   it_and_Curation_chart.png
DCS_2022_Q36_Topic_04_Repository_Features_Architecture_Choices_Large
   _Data_File_Size_chart.png
```

```
DCS_2022_Q36_Topic_04_Repository_Features_Architecture_Choices_Large
   _Data_Number_of_Files_chart.png
DCS_2022_Q36_Topic_05_Repository_Features_Architecture_Choices_Metad
   ata_chart.png
DCS_2022_Q36_Topic_06_Repository_Features_Architecture_Choices_File_
   Management_and_Reuse_chart.png
DCS_2022_Q36_Topic_07_Repository_Features_Architecture_Choices_Data_
   Privacy_chart.png
DCS_2022_Q36_Topic_08_Repository_Features_Architecture_Choices_Docum
   entation_and_Analysis_chart.png
DCS_2022_Q36_Topic_09_Repository_Features_Architecture_Choices_Prese
   rvation_chart.png
DCS_2022_Q36_Topic_10_Repository_Features_Architecture_Choices_Disco
   very_and_Reuse_chart.png
DCS_2022_Q36_Topic_11_Repository_Features_Architecture_Choices_Acces
   sibility_chart.png
DCS_2022_Q36_Topic_12_Repository_Features_Architecture_Choices_Inter
   nationalization_chart.png
DCS_2022_Q36_Topic_13_Repository_Features_Architecture_Choices_DevOp
   s_chart.png
DCS_2022_Q36_Topic_14_Repository_Features_Architecture_Choices_Scala
   bility_chart.png
DCS_2022_Q36_Topic_Summary_Repository_Features_Architecture_Choices_
   Scalability_data{.tsv,.xlsx}
DCS_2022_Q36_Topic_Summary_Repository_Features_Architecture_Choices_
   Scalability_chart.png
```

The aim of the next two questions of the survey is to find out how satisfied Dataverse installations are with the documentation and guidance provided by the Dataverse Project. Both questions are closed-ended multiple choice questions, where the respondents for different types of documentation and guidance could choose one or more of the following response options: "Needs improvement", "Neither nor", "Works fine", "Not applicable".

Question 37. Guides

In Question 37, respondents were asked to indicate how satisfied they are with the following guides which the Dataverse Project provides on their website (cf. https://guides.dataverse.org/en/latest/):

    a.   User Guide
    b.   Admin Guide
    c.   API Guide
    d.   Installation Guide
    e.   Developer Guide
    f.   Style Guide
    g.   Other

Figure 62 shows the number of submitted responses for each selected response option in Question 37.
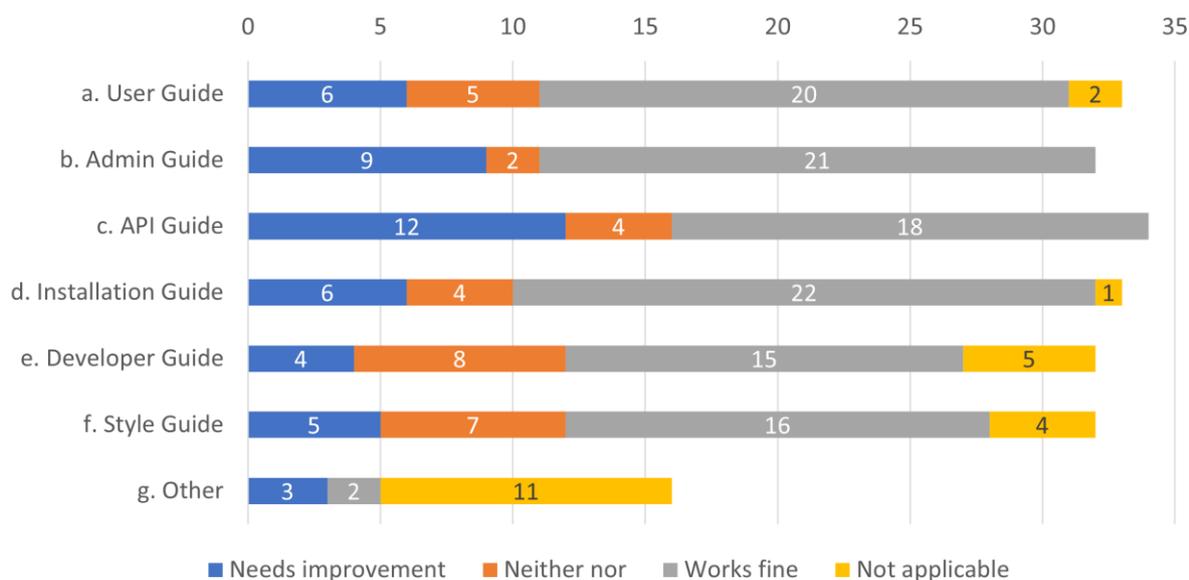
Figure 62: Results from Question 37. How satisfied is your Dataverse installation with the documentation and guidance provided by the Dataverse Project? The figure shows the number of submitted responses for each selected response option for the documentation type Guides.

All these guides work fine for a majority or a near-majority of the Dataverse installations responding to this question. The API Guide and the Admin Guide are considered to be in need for improvement by a considerable number of respondents.

For other details, the reader is referred to the supporting material.

```
### Supporting material:
DCS_2022_Q37_Guides_data{.tsv,.xlsx}
DCS_2022_Q37_Guides_chart.png
```

Question 38. Best Practices

In Question 38, respondents were asked to indicate how satisfied they are with the guidance the Dataverse Project provides on best practices for different aspects of research data archiving and publishing (cf. "Best Practices" at https://dataverse.org/), which includes the following sections:

a. Academic Credit
b. Data Citation
c. Dataverse Community Norms
d. Data Management
e. Replication Dataset Guidelines
f. Other

Figure 63 shows the number of submitted responses for each selected response option in Question 38.
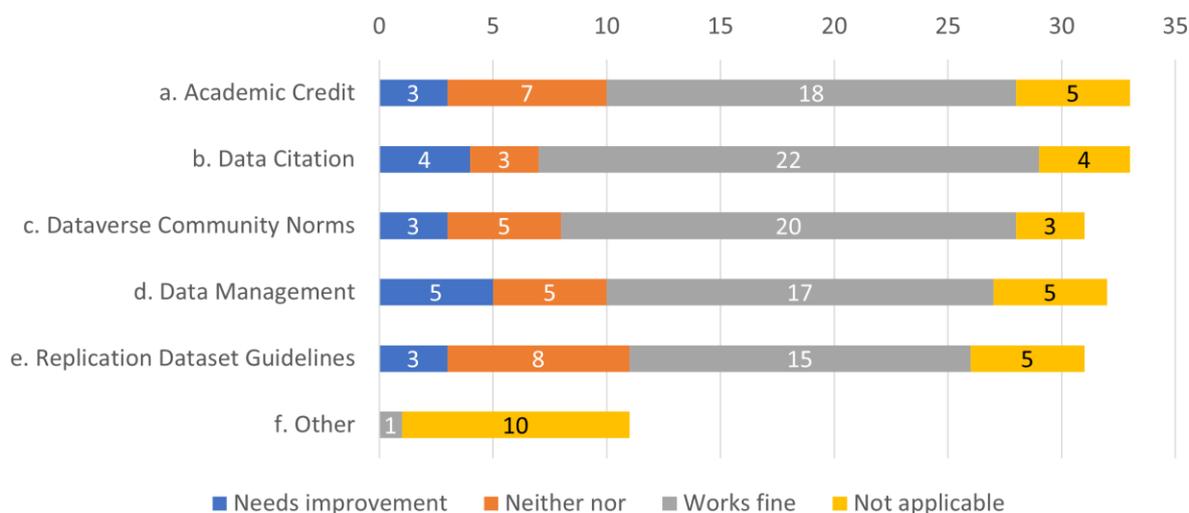
Figure 63: Results from Question 38. How satisfied is your Dataverse installation with the documentation and guidance provided by the Dataverse Project? The figure shows the number of submitted responses for each selected response option for the documentation type Best Practices.

All sections in the best-practice guidance work fine for a majority or a near-majority of the Dataverse installations responding to this question. The best-practice guidance on data management is the section with the highest numbers of votes for "needs improvement".

For other details, the reader is referred to the supporting material.

```
### Supporting material:
DCS_2022_Q38_Best_Practices_data{.tsv,.xlsx}
DCS_2022_Q38_Best_Practices_chart.png
```

Support provided / to be provided by the Dataverse community / GDCC

The aim of this section of the survey is to get an overview of how important different areas of support provided by the Dataverse community / the Global Dataverse Community Consortium (GDCC) are considered by Dataverse installations. Questions 39 to 42 are all optional closed-ended multiple choice questions, where respondents for each support area could select between the following response options: "Not applicable", "Not important", "Somewhat important", "Neither nor", "Important", "Very important". In addition to the listed support areas, respondents were given the possibility to add their own alternative(s) and/or elaborate on other matters in an open-ended / free-text field.

Question 39. Organizational Support: How important is it for your Dataverse installation that the Dataverse community (including GDCC) provides / should provide human/organizational support within the following areas?

In Question 39, respondents were asked about the importance of the Dataverse community (including GDCC) providing human/organizational support within the following areas:

a. Software installation/migration and upgrade
b. Support for install/administration (e.g., via email, issues, and/or tickets)
c. Training of DevOps and other support staff
d. Repository certification
e. PID (e.g., DOI) providing
f. Curation
g. Long-term preservation
h. Sustainable management and further development of the Dataverse software and the larger ecosystem of integrated services and tools
i. Dataverse as "archive in a box" (cf. https://github.com/IQSS/dataverse-docker)
j. Dataverse repository as a (cloud) service
k. Speakers who can represent the Dataverse community at conferences
l. Other

Figure 64 shows the number of submitted responses for each selected response option.

There is no doubt about the importance of supporting the sustainable management and further development of the Dataverse software and the larger ecosystem of integrated services and tools. 29 out of the 33 Dataverse installations rating the importance of this type of support consider it as important or very important that the Dataverse community (including GDCC) provide this kind of support.

More than half of the installations also rated the following areas of human/organizational support as important or very important:

b. Support for install/administration (25 installations)

a. Software installation/migration and upgrade (23 installations)

k. Speakers who can represent the Dataverse community at conferences (19 installations)

d. Repository certification (18 installations)

e. PID providing (18 installations)

If we include the response options "somewhat important", "important", and "very important", also the remaining areas of support reach a score of more than half of the Dataverse installations participating in the survey.
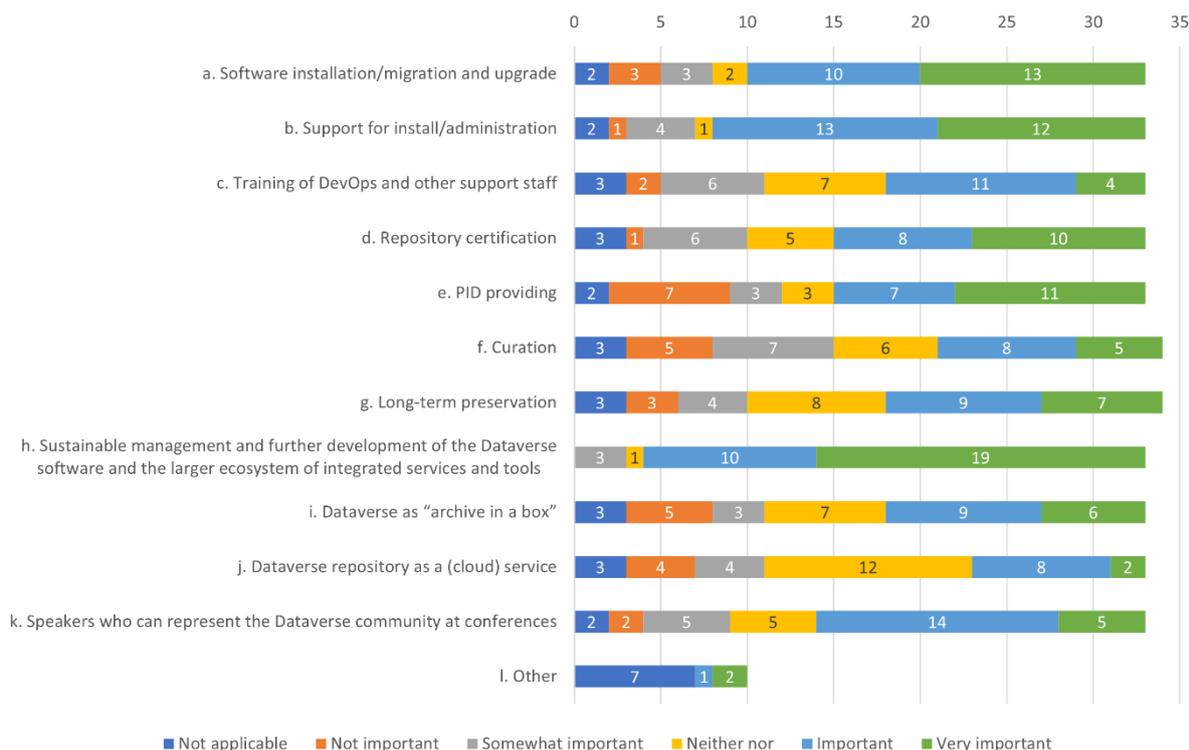
Figure 64: Results from Question 39. Organizational Support: How important is it for your Dataverse installation that the Dataverse community (including GDCC) provides / should provide human/organizational support within the following areas? The figure shows the number of submitted responses for each selected response option.

For other details, the reader is referred to the supporting material.

```
### Supporting material:
DCS_2022_Q39_Organizational_Support_data{.tsv,.xlsx}
DCS_2022_Q39_Organizational_Support_chart.png
```

Questions 40 to 42 are about sustainability and governance. They address the allocation of efforts and resources as well as well-functioning governance that are needed for developing and maintaining a community-driven open-source software.

Question 40. Sustainability: How important does your Dataverse installation consider the following existing and potential sources to ensure sustainable development and maintenance of the Dataverse software and the ecosystem of associated services and tools?

In Question 40, respondents were asked to rate the importance of the following existing and potential sources to ensure sustainable development and maintenance of the Dataverse software and the ecosystem of associated services and tools:

a. Commitment by Harvard University
b. Community commitment in the form of funding (including institutional funding and external funding)
c. Community commitment in the form of code contribution and other efforts

d. Commitment channeled through the Global Dataverse Community Consortium (GDCC)
e. Funding from The Global Sustainability Coalition for Open Science Services (SCOSS)
f. Other

Figure 65 shows the number of submitted responses for each selected response option in Question 40.
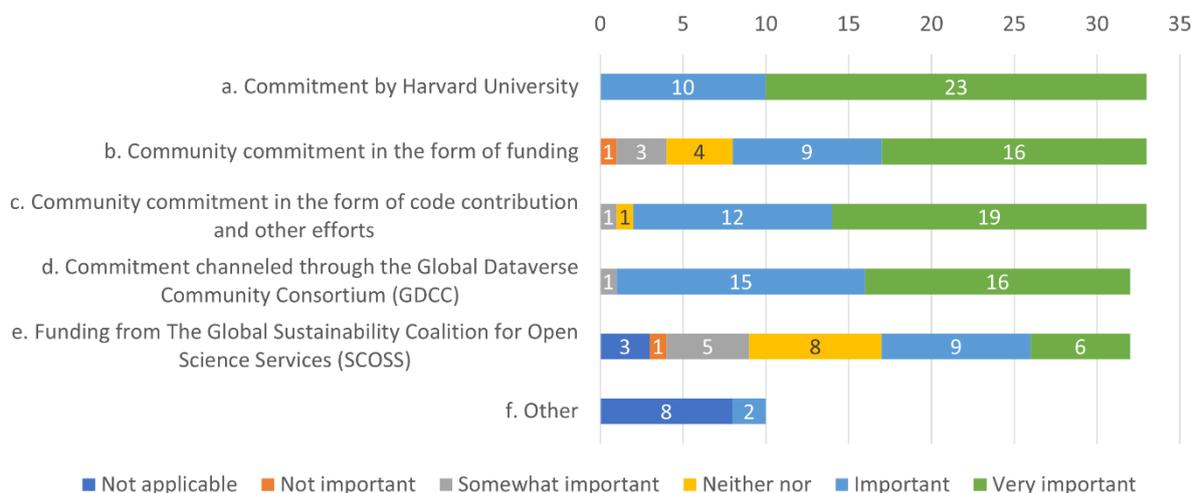

Figure 65: Results from Question 40. Sustainability: How important does your Dataverse installation consider the following existing and potential sources to ensure sustainable development and maintenance of the Dataverse software and the ecosystem of associated services and tools? The figure shows the number of submitted responses for each selected response option.

Except for funding from The Global Sustainability Coalition for Open Science Services (SCOSS), a large majority of the Dataverse installations participating in the survey consider the listed existing and potential sources to be important or very important to ensure sustainable development and maintenance of the Dataverse software and the ecosystem of associated services and tools. The commitment by Harvard University is rated as important or very important by all installations responding to this part of Question 40. Apart from one installation, the same is true for commitment channeled through the Global Dataverse Community Consortium (GDCC).

For other details, the reader is referred to the supporting material.

```
### Supporting material:
DCS_2022_Q40_Sustainability_data{.tsv,.xlsx}
DCS_2022_Q40_Sustainability_chart.png
```

Question 41. Governance: How important does your Dataverse installation consider the following existing and potential governance support for the sustainable development and maintenance of the Dataverse software and the ecosystem of associated services and tools?

In Question 41, respondents were asked to rate the importance of existing and potential means of governance support for the sustainable development and maintenance of the Dataverse software and the ecosystem of associated services and tools.

Figure 66 shows the number of submitted responses for each selected response option in Question 41.
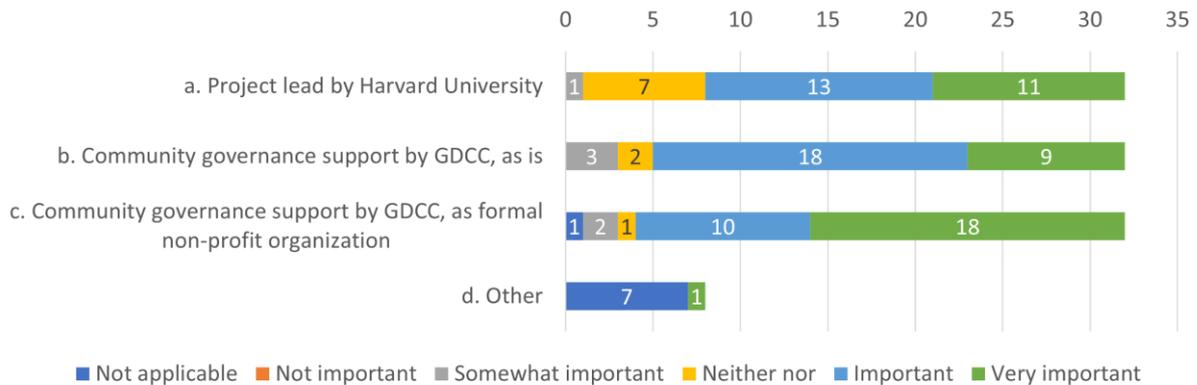

Figure 66: Results from Question 41. Governance: How important does your Dataverse installation consider the following existing and potential governance support for the sustainable development and maintenance of the Dataverse software and the ecosystem of associated services and tools? The figure shows the number of submitted responses for each selected response option.

Of the three pre-defined options, community governance support by GDCC, as a formal non-profit organization, is the means of governance support which is rated as important or very important by most of the Dataverse installations participating in the survey. Community governance support by GDCC, as is, and project lead by Harvard University are also rated to be important or very important by a large majority of the installations.

In a comment added in the free-text field, one of the respondents calls for more transparency in GDCC governance.

For other details, the reader is referred to the supporting material.

```
### Supporting material:
DCS_2022_Q41_Governance_data{.tsv,.xlsx}
DCS_2022_Q41_Governance_chart.png
```

Question 42. Governance Support by GDCC: How important is it for your Dataverse installation that the GDCC provides / should provide the following existing and potential types of governance support for the sustainable development and maintenance of the Dataverse software and the ecosystem of associated services and tools?

Question 42 is focusing on the role of GDCC for the governance of the Dataverse community. In this question, respondents were asked to indicate the importance of GDCC providing existing and potential types of governance support for the sustainable development and maintenance of the Dataverse software and the ecosystem of associated services and tools.

Figure 67 shows the number of submitted responses for each selected response option in Question 42.
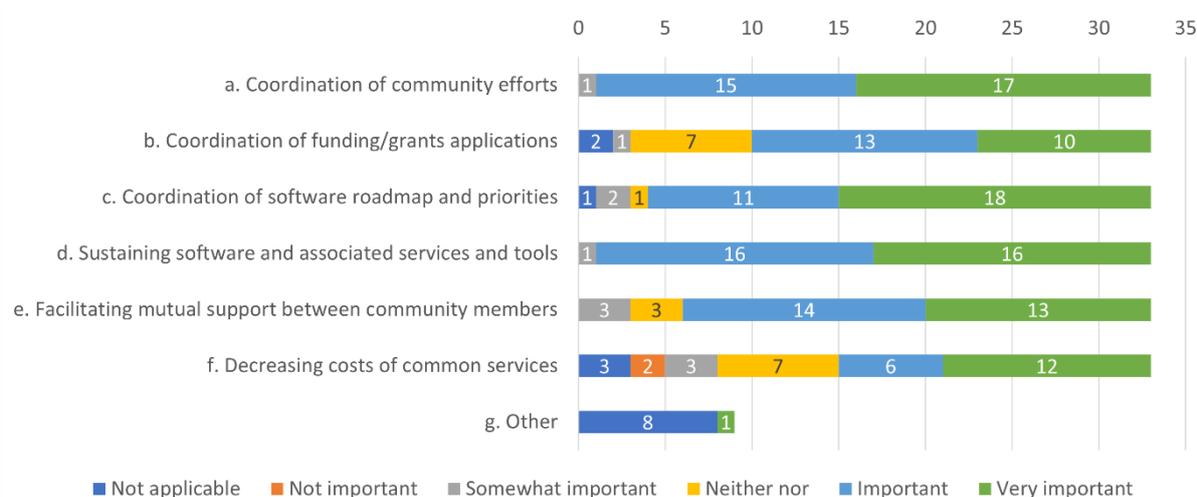

Figure 67: Results from Question 42. Governance Support by GDCC: How important is it for your Dataverse installation that the GDCC provides / should provide the following existing and potential types of governance support for the sustainable development and maintenance of the Dataverse software and the ecosystem of associated services and tools? The figure shows the number of submitted responses for each selected response option.

A large majority of the Dataverse installations participating in the survey consider it as important or very important that GDCC provide the listed types of governance support, the two most important areas being 1) coordination of community efforts and 2) sustaining the Dataverse software and associated services and tools. The majority of respondents is smaller for the importance of GDCC to provide support for decreasing costs of common services (e.g., DataCite fees).

For other details, the reader is referred to the supporting material.

```
### Supporting material:
DCS_2022_Q42_Governance_Support_by_GDCC_data{.tsv,.xlsx}
DCS_2022_Q42_Governance_Support_by_GDCC_chart.png
```

PART 5: OTHER FEEDBACK

Question 43. Other Feedback: If you want to provide any other feedback on this Survey, please add your comments or questions below (note: do not enter any personal information).

In Questions 43, which is an optional open-ended question and the only question in the last part of the survey, respondents were given the possibility to provide any other feedback on the survey. For provided feedback, the reader is referred to the anonymized raw data file in the supporting dataset.

Reference List

Arp, L. G., & Forbes, M. (2022). *It Takes a Village: Open Source Software Sustainability. A Guidebook for Programs Serving Cultural and Scientific Heritage* (February 2018; Minor Updates March 2022). LYRASIS. https://www.lyrasis.org/programs/Documents/ITAV_Guidebook_042522FINAL.pdf

Conzett, P. (2020). DataverseNO: A National, Generic Repository and its Contribution to the Increased FAIRness of Data from the Long Tail of Research. *Ravnetrykk*, *39*, Article 39. https://doi.org/10.7557/15.5514

Conzett, P. (2022). *Dataverse Community Survey 2022 – Data* (Version 1) [Dataset]. DataverseNO. https://doi.org/10.18710/UOC8CP

Conzett, P., Crabtree, J., & Myers, J. (2022). *Updates from the Global Dataverse Community Consortium (GDCC). Presentation at the Dataverse Community Meeting 2022*. https://hdl.handle.net/10037/26038

Crosas, M. (2020). Fair Principles and Beyond: Implementation in Dataverse. *Septentrio Conference Series*, *2*, Article 2. https://doi.org/10.7557/5.5334

DevOps. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=DevOps&oldid=1124816167

GDCC—The Global Dataverse Community Consortium. (n.d.). Retrieved 11 December 2022, from https://dataversecommunity.global/

IQSS/dataverse/releases. (n.d.). IQSS/Dataverse. Retrieved 8 May 2022, from https://github.com/IQSS/dataverse/releases

King, G. (2007). An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods & Research*, *36*(2), 173–199. https://doi.org/10.1177/0049124107306660

Sauro, J. (2018, September 5). 4 Classes of Survey Questions. *MeasuringU*. https://measuringu.com/survey-question-classes/

Schlatter, T., & Ji, J. (2021). *Personas for software? How and why we created archetypes for installation of an open source product* [Poster]. The information architecture conference (IAC21), Virtual. https://drive.google.com/file/d/1SA2W7MKMRXTAzFrZmjVYM-E6o9tT1OQm/view?usp=sharing

The Dataverse Project. (n.d.). Retrieved 11 December 2022, from https://dataverse.org/

Working Party of National Experts on Science and Technology Indicators. (2007). *Revised Field of Science and Technology (FOS) Classification in the Frascati Manual*. Organisation for Economic Co-operation and Development. https://www.oecd.org/science/inno/38235147.pdf