National Czech Programme | eosc

M5.1.2 | Profile-guided deposition: concept and demonstration

Document Information

Title	**Profile-guided deposition: concept and demonstration**
Project Registration Number	CZ.02.01.01/00/23_014/0008787
KA Number	5.1
Outcome Name and Number	M5.1.2 – Profile-guided deposition: concept and demonstration
Responsible Person	Ilona Trtíková, ilona.trtikova@cesnet.cz
Licence	CC BY 4.0 — https://creativecommons.org/licenses/by/4.0/

Version History

Version	Date	Comment
1.0	30.09.2026	Initial version of the document

Co-funded by the European Union

Ministry of Education, Youth and Sports of the Czech Republic

National Repository Platform for Research Data (NRP)
CESNET, zájmové sdružení právnických osob
Mailing Address: Gen. Píky 430/26, 160 00 Prague
info@eosc.cz; www.eosc.cz

Authors

Institution	Name and Surname	ORCID	Email
CEITEC, Masaryk University	Tomáš Raček	0000-0002-0296-2452	tomas.racek@muni.cz

List of abbreviations

FAIR	Findable, Accessible, Interoperable, Reusable
FIP	FAIR Implementation Profile
NRP	National Repository Platform for Research Data
API	Application Programming Interface
PDF	Portable Document Format

Table of Contents

List of Tables 7
1 Introduction 9
2 How Profile-Guided Deposition Works 10
  2.1 Shared Semantic Definitions 10
  2.2 Deposition Profiles and Guidance 10
3 Using Profile-Guided Deposition 12
  3.1 What the Builder Specifies 12
  3.2 What the Depositor Sees 12
4 Possible Uses of Annotated Files 15
5 Contribution to FAIRness 16
6 Conclusion 17
Related Outputs and Resources 18
Glossary 19

List of Figures

Figure 1: Profile-guided deposition workflow from community requirements to annotated files. 10
Figure 2: Requirements for the same transcript and guide under two profiles. 13
Figure 3 File meanings after adding the recording 14

List of Tables

Table 1: File requirements for the two illustrative interview profiles..... 12

Executive Summary

This document presents the concept and demonstration of profile-guided deposition, an approach designed to support the implementation of FAIR principles in research data repositories. It focuses on translating community-specific requirements into deposition profiles that guide users in preparing and describing files within a dataset.

The proposed approach combines shared semantic definitions with deposition profiles specifying required, recommended, and optional file types. During deposition, users receive guidance on expected files, review suggested semantic annotations, and receive feedback on whether the selected profile's requirements are met. These annotations are retained as part of the dataset metadata, supporting machine-readable descriptions of individual files.

The concept was demonstrated in a test repository based on CESNET Invenio, using illustrative profiles for recorded and transcribed interviews. The demonstration shows how different profiles can reuse the same semantic definitions while applying different file requirements.

Profile-guided deposition contributes to FAIRness by improving the consistency and machine-readability of file descriptions, supporting interoperability and facilitating future discovery and reuse of research data. The document also outlines potential extensions, including file-based filtering, dataset previews, and metadata extraction.

1 Introduction

The FAIR principles aim to make data Findable, Accessible, Interoperable, and Reusable. What this means in practice depends on the scientific domain, so each research community needs to agree how to apply these principles to its data. A FAIR Implementation Profile (FIP) is one way to record the standards and resources a community chooses.

Repositories can help depositors follow their community's choices by explaining which files to include and how to describe their meaning. This document describes how a repository can provide that guidance. A deposition profile turns those choices into file requirements and guidance for preparing a deposit, using shared terms for file meanings. Communities can agree these choices directly or draw on a formal FIP. The repository's metadata profile describes the dataset as a whole, including the study and its creators; a deposition profile adds requirements for the files the dataset contains. The repository handles access, licensing, and preservation through its other services.

The NRP analysis of procedures for implementing the FAIR principles recommends that repositories support the data standards of each research field, mainly on the basis of FIPs, and offer tools in the user interface for checking whether uploaded data comply. Profile-guided deposition provides this support for the files in a deposit.

The approach builds on established practices for describing research packages through profiles and guiding metadata entry through community-defined templates, applying these ideas to file annotation during deposition.

2 How Profile-Guided Deposition Works

Three actors take part in profile-guided deposition. The repository builder turns community requirements into profiles, shared terminology, and guidance. The depositor adds files and reviews their annotations. Consumers are people or software that use the files and their annotations to understand or process the dataset. Figure 1 shows how their work connects.

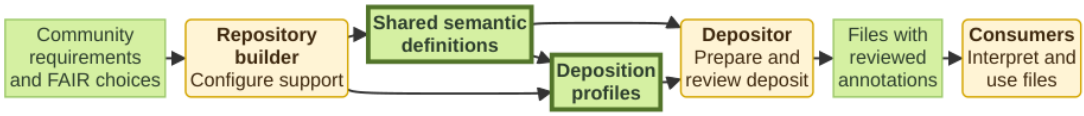


Figure 1: Profile-guided deposition workflow from community requirements to annotated files.

Profile-guided deposition connects community requirements, repository configuration, depositor review and subsequent use of annotated files.

Green rectangles show information assets; yellow rounded boxes show activities, with the responsible roles in bold. The darker borders mark two kinds of repository configuration: semantic definitions that are reused across profiles and deposits, and deposition profiles that use those definitions to specify expected files and guide the depositor.

2.1 Shared Semantic Definitions

Shared semantic definitions explain what files mean. Each definition gives a semantic type a name, a short explanation, and a stable identifier that software can use. A semantic type describes what a file represents within the dataset, independently of its format: an interview transcript remains a transcript whether it is saved as plain text or PDF.

The repository defines these meanings separately from its profiles, so each profile can refer to the same definitions. For example, profiles for recorded and transcribed interviews can both use the semantic type interview-guide for the questions or prompts used by the interviewer. File annotations apply these meanings to individual files. A file can have more than one semantic type when it contains several kinds of information.

2.2 Deposition Profiles and Guidance

A deposition profile is a repository configuration that describes which files a particular kind of dataset should contain. Using the shared semantic definitions, it identifies required, recommended, and optional kinds of files, any limits on their numbers, and guidance for the depositor. A recordedinterview profile, for example, may require at least one interview recording and allow an interview transcript.

The repository uses this configuration in its deposition interface: it explains which files are expected and can suggest semantic types from filenames. The depositor reviews these suggestions and can

correct them. Using the reviewed types and uploaded files, the repository checks whether the expected files are present in the specified numbers. It can then give concrete advice, such as asking the depositor to add a missing interview recording.

These checks look at which files are present and how they are described. Checking whether the data themselves are scientifically sound is a separate, domain-specific task. The reviewed file meanings are saved with the files, so people and software can use them later.

3 Using Profile-Guided Deposition

The following example illustrates the concept: a depositor prepares an interview recording, its transcript, and an interview guide. Switching between recorded-interview and transcribed-interview profiles shows how the same file meanings can be used with different expectations about which files to include.

Note: The demonstration ran in a test repository built on CESNET Invenio, the Invenio-based repository platform used within NRP. It used the stable OARepo 14.2 release of the platform (oarepo-app 7.0, based on InvenioRDM 14). Profile support was added as an extension (https://github.com/sbncbr/profile-guided-deposition/tree/m5.1.2): the configuration, metadata fields, controls in the deposit form's Files section, and a publication check. Uploading files and saving drafts work as in any CESNET Invenio repository.

3.1 What the Builder Specifies

The builder defines semantic types for the files and selects which files each profile requires. For example, interview-recording identifies a recording of an interview, interview-transcript identifies its written transcription, and interview-guide identifies the questions or prompts used by the interviewer.

The two profiles in Table 1 are simple examples for this demonstration. They use the same semantic definitions and set different file requirements.

Table 1: File requirements for the two illustrative interview profiles

Profile	**Required file**	**Optional supporting files**
Recorded interview	At least one interview recording.	Interview transcript and interview guide.
Transcribed interview	At least one interview transcript.	Interview guide.

Both profiles use the same definition of interview-guide. Their requirements differ: the recorded-interview profile requires a recording, while the transcribed-interview profile requires a transcript. The semantic definitions explain what the files represent; the profiles specify which files are expected.

A PDF could be an interview transcript or an interview guide. The interface can suggest an assignment from a filename such as interview-guide.pdf. The depositor reviews the suggestion and can correct it.

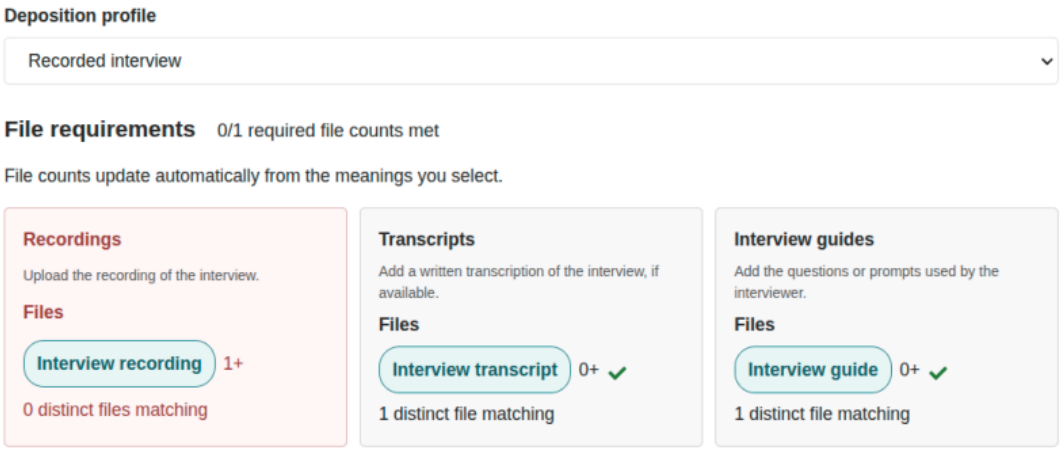
3.2 What the Depositor Sees

The depositor selects a profile, uploads files, and reviews their suggested meanings. This example uses synthetic files representing an interview recording, its transcript, and the guide listing the interview questions.

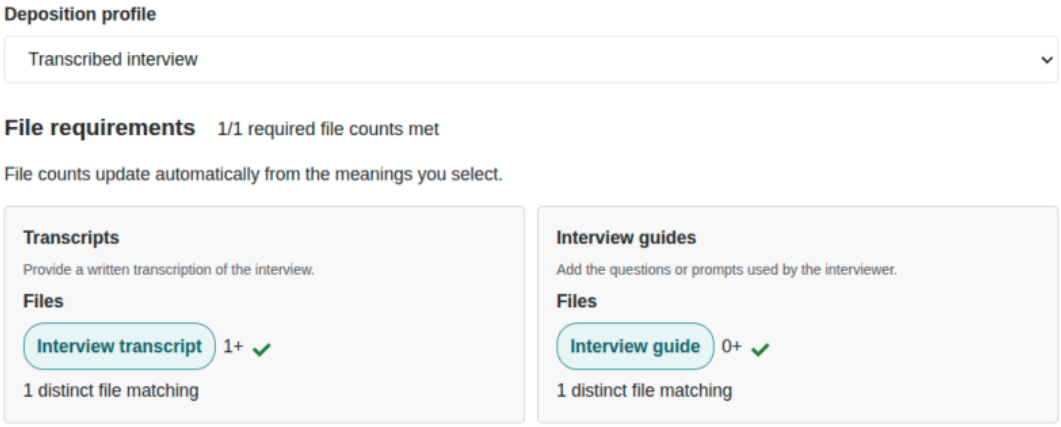
First, the depositor selects Recorded interview and uploads interview.txt and interviewguide.pdf. Figure 2(a) shows that the transcript and guide are present, but the required recording is missing. Each card explains which files are expected and shows their

semantic-type tags and counts. A green check indicates that the file-count requirement is met. The depositor can save the draft while preparing the remaining files.

Selecting Transcribed interview changes the requirements: the uploaded transcript now satisfies the required file count, as shown in Figure 2(b). The Interview guide has the same meaning in both profiles. The transcript and guide keep their annotations.



(a) *Recorded interview.*



(b) *Transcribed interview.*

Figure 2: Requirements for the same transcript and guide under two profiles. (a) Recorded interview

Panel (a) shows that the Recorded interview profile still requires a recording, while panel (b) shows that the Transcribed interview profile has its required transcript. Both profiles allow an interview guide using the same semantic type.

Next, the depositor returns to **Recorded interview** and uploads interview.wav. Figure 3 shows the selected meanings: **Interview recording**, **Interview transcript**, and **Interview guide**. The depositor reviews these labels, corrects them if needed, and can add a description to each file. The recording requirement is now satisfied.

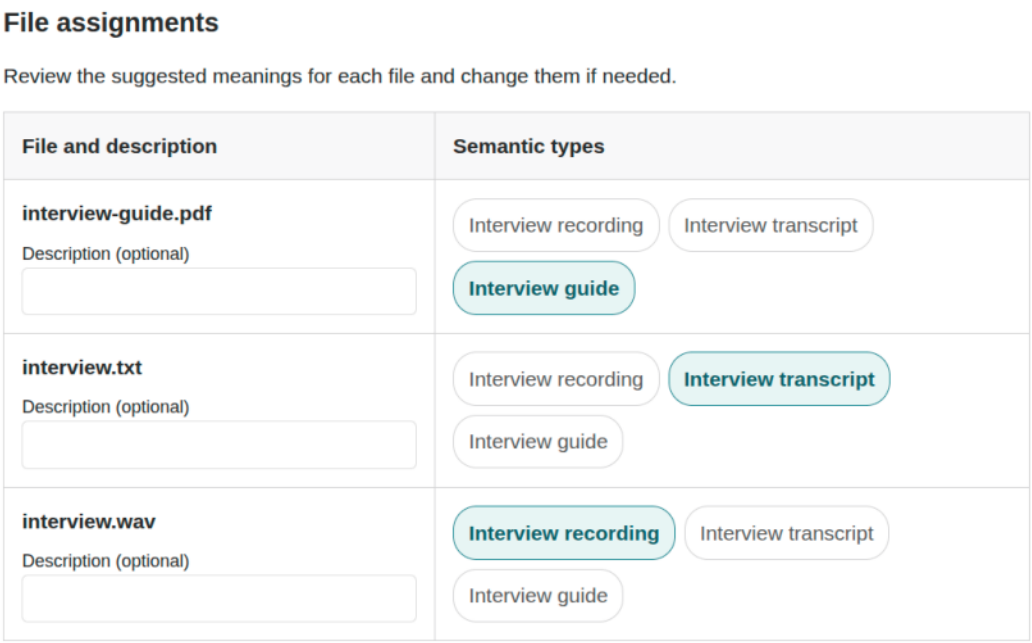


Figure 3 File meanings after adding the recording

Selected tags use the same labels as the requirement.

The draft now meets the profile's file requirements: it contains the required recording, and thetranscript and guide are annotated as optional supporting files. A recording alone would also meet these requirements. The depositor then describes the study itself, such as its title and creators, in the repository's other form sections.

4 Possible Uses of Annotated Files

Reviewed file meanings are saved alongside the uploaded files. A repository can use these annotations to help readers find, explore, and use interview data. The examples below describe possible extensions.

Browsing and filtering. Filters such as “Contains interview recordings” or “Contains transcripts” help readers find datasets with the material they need, across different deposition profiles.

Dataset views and previews. A dataset page can organise recordings, transcripts, and guides, with an audio player or document preview for supported formats. Readers can listen, read, or select files for download.

Metadata extraction. A service can read a recording’s duration and display it beside the audio player, helping readers estimate listening time. The repository can store such values together with a reference to the file they came from.

Note: Other domains have their own uses, such as extracting acquisition settings from microscopy images or simulation properties from molecular-dynamics files.

5 Contribution to FAIRness

Profile-guided deposition supports FAIRness by making community file requirements explicit and retaining file meanings as machine-readable annotations. These annotations describe what individual files represent within the dataset, helping software identify candidate inputs for automated processing. When indexed for discovery, they can help readers find datasets containing the material they need. Consistent semantic identifiers with accessible shared definitions let people and software interpret files in the same way across the repository's deposits. The example uses repositorylocal terms; adopting established community vocabulary terms can extend this interoperability to other repositories and tools. Descriptions of what each file represents make the files easier to reuse.

Annotations are saved with the record metadata, so the repository API can provide them to people and software with access to the record. Other metadata exports, such as DataCite, need an explicit mapping to carry them.

Profiles also help depositors include the files their community expects. This adds to what the repository already provides: identifiers, descriptive metadata, access conditions, licensing, and preservation.

6 Conclusion

This document presented the concept and demonstration of profile-guided deposition as an approach to supporting community-specific requirements in research data repositories. By combining shared semantic definitions with deposition profiles, the approach provides guidance on expected files and enables depositors to review file annotations and check whether profile requirements are met.

The concept was demonstrated in a test repository based on CESNET Invenio, using illustrative profiles for recorded and transcribed interviews. The demonstration showed how different profiles can reuse shared semantic definitions while specifying different file requirements.

Profile-guided deposition contributes to FAIRness by supporting consistent, machine-readable descriptions of individual files and helping depositors prepare datasets according to community expectations. Further possibilities include improved discovery, dataset previews, and metadata extraction. Extending interoperability beyond individual repositories would require the adoption of established community vocabularies and appropriate metadata mappings.

Related Outputs and Resources

- Trtíková, I. et al. Analýza a návrh postupů k naplňování FAIR principů v rámci NRP/NDI. Národní repozitářová platforma pro výzkumná data / CESNET, version 1.2 (2026). https://doi.org/10.71495/5dv6-v116.
- Profile-Guided Deposition – Demonstration Implementation. Source code and configuration developed for milestone M5.1.2. https://github.com/sb-ncbr/profile-guided-deposition/tree/m5.1.2

Glossary

Term	Definition
CESNET Invenio	The Invenio-based repository platform used within NRP, distributed as the OARepo software packages.
Deposition profile	File requirements and depositor guidance for a kind of dataset, using shared semantic types. It is a specific kind of application profile and complements the repository's metadata profile, which describes the dataset as a whole.
FAIR	Findable, Accessible, Interoperable, and Reusable: principles for making data useful to people and software.
FAIR Implementation Profile (FIP)	A community's recorded choices of standards and resources for implementing FAIR.
File annotation	Metadata recording a file's reviewed semantic types and optional description.
NRP	National Repository Platform for Research Data: shared infrastructure and services for Czech research-data repositories.
Semantic type	A file's meaning within a dataset, identified by a stable term independently of format, such as Interview transcript.