# Automating GDPR-Compliant Consent Forms: A Modular and Interdisciplinary Approach

Maryam Mohammadi[1,*], Aliena Strathmann[1], Katja Politt[2,1], Annett Jorschick[1] and Hendrik Buschmeier[1]

[1]CRC 1646 'Linguistic Creativity in Communication', Bielefeld University, Bielefeld, Germany
[2]Rostock University, Rostock, Germany

**Abstract**

Legal and ethical requirements surrounding the processing of (personal) data have become increasingly complex, introducing a great degree of overhead to scientific research requiring the processing of data generated by human participants. In order to facilitate the process of ethical and GDPR-compliant data processing, we introduce a modular, ontology-based framework and first implementation of an automation tool that balances technical, legal and user group requirements based on interdisciplinary expertise; allowing researchers to easily generate tailored, comprehensible data processing information and consent forms for study participants. A future goal is to expand this framework into a more extensive data management platform, to promote ethical and legally compliant use and re-use of (linguistic) study datasets across research teams and domains according to Open Science principles.

**Keywords**

consent forms, personal data, GDPR, modules, ontology

## 1. Introduction

Research in linguistics inherently involves data from humans, which often counts as personal or sensitive data. This can limit Open Science practices and adherence to the FAIR principles [1], such as data publication, sharing, and reuse. These limitations are not a late-stage documentation issue; they shape study design, the information provided to participants, and the range of permitted downstream operations on collected data. As a result, researchers face recurring challenges: legal and ethical requirements must be built into the planning stage, yet the resulting decisions are often ad hoc and poorly connected to downstream data handling. Recent work has therefore argued for research data management infrastructures that integrate automated support for planning, collection, storage, use, reuse, and sharing of data while explicitly accounting for ethical and legal constraints, thereby promoting Open Science practices [e.g., 2]. Building on this work, we present RUDI ("Research-centered User-oriented Data Infrastructure"), an infrastructure developed within the "INF" sub-project of the Collaborative Research Center CRC 1646 *Linguistic Creativity in Communication* at Bielefeld University, Germany, where, across projects, heterogeneous study designs, data types, and participant populations are the norm rather than the exception.

In this paper, we describe a framework for addressing these challenges. The framework encompasses three perspectives: (i) a *user* perspective that specifies study-specific intents and constraints, (ii) a *legal* perspective that contributes normative requirements, obligations, and permissible conditions under GDPR [3] and ethics, and (iii) a *technical* perspective that represents the former two formally as legal and domain properties of the study, making it suitable for validation, traceability, and downstream processing. These three perspectives are the basis of a three-component platform, with each component corresponding to one perspective. We report an initial implementation that operationalizes a Consent

*4th Privacy & Personal Data Management Session of the Solid Symposium 2026*
*Corresponding author.
maryam.mohammadi@uni-bielefeld.de (M. Mohammadi); aliena.strathmann@uni-bielefeld.de (A. Strathmann); katja.politt@uni-bielefeld.de (K. Politt); annett.jorschick@uni-bielefeld.de (A. Jorschick); hbuschme@uni-bielefeld.de (H. Buschmeier)
0009-0007-8747-029X (M. Mohammadi); 0000-0002-4912-2653 (K. Politt); 0009-0004-0776-7113 (A. Jorschick); 0000-0002-9613-5713 (H. Buschmeier)

Form Wizard together with a corresponding ontology that specifies the necessary vocabulary and underlying taxonomy. The wizard allows researchers to record properties of a study as structured metadata and generate clear, legally and ethically compliant documents specific to the study, mainly addressing perspectives (i) and (ii). This, in turn, informs subsequent decisions about data handling: for any dataset, or even 'data point' [2], the recorded study properties and participants' individual consent choices can be used to determine which operations (e.g., processing, sharing, publication, or controlled access) are permissible.

Questionnaire-driven tools already demonstrate how parts of this process can be standardized. The 'DARIAH Consent Form Wizard' [4] supports template-based generation of consent forms, and the 'Ethiktool' [5, 6] provides software-guided collection of information relevant for ethics review while generating participant information and privacy-related documents. These systems motivate guided standardization, but they typically treat the generated documents and application materials as the final output rather than as inputs to downstream, machine-processable decisions about permissible data operations, addressing perspective (iii).

The wizard-based *technical* application advances these earlier implementations by replacing static, manually curated templates with dynamically generated configurations that adapt expert-provided *legal* foundations to the concrete needs of *users* (which include both researchers and study participants). In parallel, the platform's ontology layer builds on established standards (such as the GDPR-compliant Data Privacy Vocabulary; [7]) to ensure semantic coherence, improve metadata accessibility, and support interoperability in subsequent data management and reuse workflows. Together, the wizard and the ontology provide a systematic, maintainable, yet dynamic approach to the aforementioned challenges of linguistic data management by integrating legal and technical requirements into a unified framework.

The paper is structured as follows: In Section 2, we introduce the modular approach of the dynamic framework and describe the interfaces between each of the components. Section 3 presents the framework architecture, including its ontological foundation (§3.1) and the implementation of the consent form wizard (§3.2). Section 4 summarizes the project and outlines directions for future development.

## 2. Modular approach

Linguistic research data is the initial application domain of the platform, where primary use cases are drawn from. Linguistics is an interdisciplinary science, including e.g., psychology, sociology, computational and clinical linguistics, which raises two main challenges for a modular approach: variability in data types and formats, and diversity in participant populations. First, linguistic data types vary widely across different dimensions, including modality (e.g., text, audio, video, reaction times) and research setting (e.g., corpora, fieldwork, experiments). While not all linguistic data inherently qualify as sensitive or personal data per se, associated demographic and contextual metadata typically constitute personal data and therefore require careful governance. This diversity limits the applicability of static text modules for consent forms, ethics applications, and data management plans.

Second, linguistic studies often involve diverse participant groups. Apart from neurotypical adults, research may target second-language learners, children, or individuals with mental or physical impairments, some requiring guardian involvement [8]. Furthermore, legal texts are often lengthy and complex, which hinders comprehension not only for some participant groups, but also for researchers without legal training. Following GDPR principles, participants must be informed in clear and accessible language; thus, researchers must understand consent requirements to communicate them accurately.

To ensure that these challenges are reflected in the framework, we adopt an interdisciplinary approach in which domain expertise (linguistics), normative requirements (legal/ethics), and technical implementation (computation) are treated as separate but coordinated inputs. The framework relies on configurable templates and a modular architecture to capture expert knowledge through one-off configuration and accommodate evolving legal and domain-specific requirements. The resulting model is structured into three interacting modules: *user, legal, technical* and the remainder of this section specifies how information flows across their pairwise interfaces.

## 2.1. The User–Legal interface

In our framework, *user* refers to two stakeholder groups: (i) *researchers*, who design studies and collect, use, and manage data, and (ii) study *participants*, who provide data ('data-subjects' in GDPR). The user–legal interface reflects the fact that both groups must be adequately informed about the contents and implications of consent, albeit in different ways. Researchers need to know which study properties relate to which legal and ethical obligations. Many studies can be conducted without collecting unnecessary personal data, adhering to the principle of data minimization. Providing researchers with accessible information of what counts as personal or sensitive data and why these might not be necessary to the research design helps prevent superfluous data collection. Additionally, it ensures that planned data-handling and sharing practices remain legally feasible. Participants, in turn, require information that allows informed and voluntary consent. In practice, consent forms are often signed without full comprehension due to their complexity, despite the legal requirement that information must be provided in a manner appropriate to the participants' level of understanding.

The user–legal interface of our framework supports both groups as follows: Building on a comprehensive list of linguistic use cases (compiled by Mohammadi et al. [9]; e.g., data types, participant types, data collection purposes, etc.), legal experts in the project first identify user needs and properties, develop legal use cases and requirements, and then translate them into templates and guidelines implemented at the technical level. The result is a structured set of requirements and text templates that integrate domain-specific study properties with legal constraints. This output forms the basis for the technical operationalization in the next section.

## 2.2. The Legal–Technical interface

The operationalization of the legal requirements outlined above within the software architecture consists of two main steps: First, all legally relevant study properties are defined in a configuration template. Second, a dynamic input form is generated from a *form steps template* configuration that factors in legal constraints such as conditionally required information, and maps form entries made by the researcher to the corresponding study properties.

Using configuration templates allows specifying all possible scenarios in advance through leveraging the immediate guidance of a legal expert. Besides requesting conditionally required information from the researcher, the dynamic and modular form generation process also enables (i) displaying contextual, supplementary legal information in the form of tooltips for researchers, (ii) providing conditional recommendations and explanations for legally complex combinations of study properties, and (iii) enforcing conditionally fixed values: for instance, a form requiring guardian consent is added automatically when the researcher specifies that participants are under 14 years of age in the study properties.

At a later stage, the application will automatically distinguish between data that counts as personal/sensitive and data that does not. Presently, the application relies on the researcher's own information on whether such data is being processed, and links to a third-party online tool (iVA; [10]) to aid the decision in case of uncertainty. Future versions will integrate iVA's 4-step decision process directly into the tool's configuration templates. Once personal/sensitive data is used, an additional, more detailed privacy policy as a separate document accompanying the consent form is generated.

## 2.3. The Technical-User Interface

To ensure a smooth workflow, the technical implementation is grounded on two bases. First, we adhere to established principles of interface and web design, in particular Nielsen's ten usability heuristics [11]. Second, we employ an iterative development process that incorporates continuous user testing and systematic integration of user feedback [12]. As outlined above, the legal concepts were collected in clear and understandable language within the regulatory compliance. The wizard, as a modular and configurable framework, provides contextual help boxes and ensures that users can readily access explanations of legal complexities and domain-specific terminology.

**Workflow facilitation for researchers** Condition-based rendering of only the necessary form input steps mitigates most of the need for legal expertise, while expandable information tooltips provide optional legal explanations and clarifications. Conditional suggestions inform the user of ways to simplify participant consent and data management where applicable[1]. User inputs are immediately reflected in a live preview of the required output documents.

**Workflow facilitation for study participants** The phrasing of the generated documents is carefully chosen to focus on how aspects of the study affect the participant; details about data processing, potential risks, and participant rights are communicated clearly in a structured manner. The framework is designed to provide participants with full modular control over which aspects of data processing they consent to, ensuring that their consent is well-informed and voluntary. The participants' individual consent decisions can later be linked to all related data points.

## 3. Dynamic Framework

Based on the three perspectives outlined in Sections 1 and 2, we propose a corresponding three-layer architecture: (i) an ontology-based infrastructure providing taxonomies for required components; (ii) a wizard-based consent platform that guides researchers in generating customized consent forms as well as collecting online consents from study participants; and (iii) a data management engine for querying and sharing data in line with individualized consent (employing the ontologies to enable future integration with other platforms). The next section presents the ontology layer, followed by the wizard platform, while the data management layer is deferred to future work (see Section 4).

### 3.1. Ontology Foundation

Mohammadi et al. [9] assess the needs of (linguistic) researchers by identifying relevant data types and associated (meta)data. While such information could be hard-coded into a platform, principles of data visibility and reusability require newly collected data to be linked to existing resources. Inconsistent terminology within a domain undermines discoverability. For example, queries for data from *teenagers, adolescents*, or *young children* cannot be reliably resolved without explicit semantic links. We therefore employ ontologies as taxonomies and controlled vocabularies to align concepts semantically and connect our data to the broader semantic web network [13].

Recent work has examined personal data from multiple perspectives [see, 7, among others,]. Since linguistic demographic data largely overlap with personal data categories, we adopt and adapt established ontologies and standards. For widely used categories, we rely on ISO standards, including ISO 639-3 for language codes [14] and ISO 3166-1 for country codes [15]. For more specialized data, we use domain-specific vocabularies such as BioPortal and the WHO International Classification of Diseases (ICD) for medical information. To model personal data, we adopt the GDPR-compliant Data Privacy Vocabulary (DPV; [16, 17]).

While these standards improve interoperability and data visibility, they do not fully capture domain-specific requirements in experimental linguistics. We therefore develop the *eXperimental Linguistics* taxonomy (XLing), which defines a minimal, extensible set of field-specific terms and is designed for reuse across frameworks. XLing entries are also linked to CLARIN vocabularies for future integration. [2] We employ well-established schemata, including the Resource Description Framework Schema (RDF schema), Dublin Core Terms (dcTerms), and the Simple Knowledge Organization System (SKOS, [18]). These schemata are dynamically implemented within the platform, allowing newly added values to be immediately integrated into the application.

[1] For example, by indicating when anonymization or pseudonymization may be feasible for certain types of personal data.
[2] While XLing is still under development, it is planned for independent release as an open-source taxonomy by the end of the project, enabling further development and broader reuse.

### 3.2. Wizard Platform

In its current development stage, the wizard guides researchers through a step-by-step form with queries about the properties of their study pertaining to legal aspects of data collection as well as broader contextualization for sharing and re-use. The set of properties that a study may assume, as well as the form's input steps and evaluation rules, may be configured to accommodate different research domains. To facilitate configuration and ensure consistency, the platform supports the import of study properties and answer option values from RDF taxonomies (see 3.1).

During form completion, the wizard immediately and continuously evaluates the current values of the study's properties. It introduces additional questions and/or text parts in the output documents whenever required by applicable legal constraints. In other words: as researchers fill in their study design, the wizard automatically adapts the output documents, ensuring that follow-up information is requested only when necessary. The output documents are generated from configurable XML templates, and shown reflecting the researchers' input in live preview tabs.

Once the form is considered filled and valid by the wizard, researchers are able to download the output forms as a bundle of PDF files, ready to hand off to the participant for transparency and signing of consent to ensure legal and ethical research practice. Apart from the PDF files, researchers also have the option to export their study configuration as a JSON file which can later be re-imported, e.g., for repeated or slightly adapted follow-up studies.

While the wizard currently only handles direct interaction with researchers and thus assumes the role of a highly configurable, dynamic information and consent form generator, it will be integrated into a broader data management platform with two core capabilities: (i) Ability for participants to give full or partial consent[3] directly within the platform, and storage of consent instances in a secure database (see also Section 4), eliminating the need for manually distributing and managing (signed) legal documents. (ii) Mapping consent instances to dataset metadata, facilitating legally compliant sharing and re-use of collected data within the CRC as well as among external collaborators and colleagues in the field.

## 4. Summary and Future Perspectives

In this paper, we proposed a modular framework that integrates user, legal, and technical perspectives and aligned them with model components, each developed together with experts from the respective disciplines. Moreover, we implemented a wizard-based platform that helps researchers (who usually lack the legal expertise) in setting up studies and automatically generating legally compliant consent forms, tailored to the specific requirements of their experiment. Crucially, the platform is built on standard ontologies/taxonomies as the foundation for our (meta)data representation.

The platform is currently in a pilot phase for evaluation and iterative refinement. In the next version, we plan to implement object-based database storage for consent records, enabling researchers to register on the platform, define studies, generate consent forms, and retrieve them for editing and online signing. Signed consent forms will be persistently stored and integrated into a broader consent lifecycle workflow, including withdrawal, rights exercise, and usage control. A subsequent version will extend the system with a data management framework that links consent records to the collected data, thereby supporting controlled data sharing and long-term dataset reusability among (linguistic) researchers.

## Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): CRC 1646/1 2024 – 512393437, Project INF.

## Declaration on Generative AI

The authors used GPT-4 for grammar checks on some sentences and take responsibility for the content.

[3]Partial consent means the ability to opt out of consenting to specific processing steps, e.g., publication or third-party sharing.

## References


[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. doi:10.1038/sdata.2016.18.

[2] A. Jorschick, P. T. Schrader, H. Buschmeier, What can I do with this data point? Towards modeling legal and ethical aspects of linguistic data collection and (re-)use, in: Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024, ELRA and ICCL, Torino, Italy, 2024, pp. 47–51. URL: https://aclanthology.org/2024.legal-1.8.

[3] GDPR, Regulation (EU) 2016/679 of the European Parliament and of the Council, Official Journal of the European Union L 119 (2016) 1–88. URL: https://data.europa.eu/eli/reg/2016/679/oj.

[4] V. Hannesschläger, W. Scholger, K. Kuzman, The DARIAH ELDAH consent form wizard, in: DARIAH Annual Event 2020: Scholarly Primitives, 2020.

[5] A. Bendixen, E.-M. Berens, T. G. G. Wegner, W. Einhäuser, K. Blask, Data (re-)use in human-participant research: Guided composition of informed consent forms, in: Abstracts of the 2nd Conference on Research Data Infrastructure (CoRDI), Zenodo, 2025. doi:10.5281/zenodo.16735895.

[6] A. Bendixen, T. G. G. Wegner, W. Einhäuser, Facilitating ethics application and review for interdisciplinary human-participant research via software-based guidance and standardization, in: O. K. Bertolt Meyer, Ulrike Thomas (Ed.), Hybrid Societies: Humans Interacting with Embodied Technologies, volume 1, Springer, Chemnitz, Germany, in press.

[7] H. J. Pandit, A. Polleres, B. Bos, R. Brennan, B. Bruegger, F. J. Ekaputra, J. D. Fernández, R. Gachpaz Hamed, E. Kiesling, M. Lizar, E. Schlehahn, S. Steyskal, R. Wenning, Creating a vocabulary for data privacy: The first-year report of data privacy vocabularies and controls community group (DPVCG), in: Proceedings of the 18th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2019), Springer, Rhodes, Greece, 2019, pp. 714–730. doi:10.1007/978-3-030-33246-4_44.

[8] P. T. Schrader, M.-L. Joppek, Datenbasierte Forschung und Einwilligungsunfähige. Zulässigkeit der Verarbeitung von Daten Minderjähriger und geistig eingeschränkter Personen, Zeitschrift für Datenschutzrecht 2025 (2025) 613–618. URL: https://beck-online.beck.de/?vpath=bibdata/zeits/ZD/2025/cont/ZD.2025.613.1.htm.

[9] M. Mohammadi, K. Politt, A. Jorschick, Assessing data management and compliance in large research collaborations via knowledge bases: A semi-structured interview approach, F1000Research 15 (2026) 37. doi:10.12688/f1000research.173178.1, version 1; peer review: awaiting peer review.

[10] M. Herklotz, L. Oberländer, iVA: Ein interaktiver Virtueller Assistent von BERD@BW zur Aufbereitung von Rechtsfragen im Bereich Open Science, heiBOOKS, 2022, p. 306–313. doi:10.11588/heibooks.979.c13742.

[11] J. Nielsen, Heuristic evaluation, in: J. Nielsen, R. L. Mack (Eds.), Usability Inspection Methods, Wiley, New York, NY, USA, 1994, pp. 25–62.

[12] M. Matera, F. Rizzo, G. T. Carughi, Web usability: Principles and evaluation methods, in: E. Mendes, N. Mosley (Eds.), Web Engineering, Springer, Berlin, Germany, 2006, pp. 143–180. doi:10.1007/3-540-28218-1_5.

[13] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Scientific American 34–43 (2001).

[14] ISO, ISO 639:2023 – Code for individual languages and language groups, International Standard 639:2023, International Organization for Standardization (ISO), Geneva, Switzerland, 2023. URL: https://www.iso.org/standard/74575.html.

[15] ISO/IEC, ISO/IEC 3166-1:2020 – Codes for the representation of names of countries and their subdivisions – Part 1: Country code, Standard 3166-1:2020, International Organization for Standardization (ISO), Geneva, Switzerland, 2020. URL: https://www.iso.org/standard/72482.html.

[16] H. J. Pandit, B. Esteves, G. P. Krog, D. Golpayegani, J. Flake, Data Privacy Vocabulary (DPV) – version 2.0, in: The Semantic Web – ISWC 2024, Springer, Cham, Switzerland, 2025, pp. 171–193.

doi:10.1007/978-3-031-77847-6_10.

[17] B. Esteves, D. Golpayegani, G. P. Krog, H. J. Pandit, J. Flake, P. Ryan, Data Privacy Vocabulary (DPV), version 2.2, Final Community Group Report, World Wide Web Consortium, Wakefield, MA, USA, 2025. URL: https://w3c.github.io/dpv/2.2/dpv/.

[18] A. Miles, S. Bechhofer, SKOS Simple Knowledge Organization System Reference, W3C Recommendation, World Wide Web Consortium, Wakefield, MA, USA, 2009. URL: http://www.w3.org/TR/skos-reference.