



SP-EU

**Social Prescribing to promote and improve access to health and care services
for people in vulnerable situations in Europe**

Horizon Europe – 101155873

D6.1 – ALTAI-Report

WP leader	Dr. Michael Sperber (ACT)
Author(s)	Raoul Schlotterbeck (ACT) Markus Schlegel (ACT)
Version	V1
Due date of delivery	30.06.2025
Actual date of delivery	25.06.2025
Dissemination level	PU

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101155873.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authorities can be held responsible for them.

Table of Contents

- Table of Contents 2**
- Abbreviations 3**
- Executive Summary 4**
- 1. Introduction 5**
- 2. Assessment 5**
 - 2.1 Human Agency and Oversight 5
 - 2.2 Technical Robustness and Safety 6
 - 2.3 Privacy and Data Governance 6
 - 2.4 Transparency 7
 - 2.5 Diversity, Non-discrimination and Fairness..... 8
 - 2.6 Societal and Environmental Well-being 8
 - 2.7 Accountability..... 9
- 3. Conclusion 9**

Abbreviations

AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
EU	European Union
FAQ	Frequently Asked Question
LGBTIQ	Lesbian Gay Bisexual Trans Inter Queer
RDF	Randomized Control Trial
SHACL	Shapes Constraint Language
SP	Social Prescribing
UI	User Interface

Executive Summary

This report presents an assessment of a link worker resource management application with AI-assisted RDF formatting. The application aims to help link workers manage contacts and resources for social prescribing, particularly for three target groups: elderly individuals, first-generation immigrants, and members of the LGBTQ community across EU countries. The AI component assists users in formatting free-text entries into structured RDF data, powered by an Ollama server currently running the Phi-4 model.

1. Introduction

This report presents an assessment of a link worker resource management application with AI-assisted RDF formatting. The report is based on the Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment and will be updated regularly.

The application assessed herein aims to help link workers manage contacts and resources for social prescribing, particularly for three target groups: elderly individuals, first-generation immigrants, and members of the LGBTQ community across EU countries. It contains two AI components: one assists users in formatting free-text entries into structured RDF data, powered by an Ollama server currently running the Phi-4 model; the other one performs semantic area searches, allowing users to find relevant services geographically by matching natural language queries with OpenStreetMap data using the BGE-M3 embedding model.

2. Assessment

2.1 Human Agency and Oversight

Both AI tools operate under explicit user control and require manual activation through clearly marked interface elements.

One AI system is designed to assist users in converting free-text input into RDF data using a controlled vocabulary. Activation of the AI tool is entirely optional and triggered manually by the user via a clearly marked button. Once the AI generates a suggestion, users can choose to discard it or embed it into their RDF graph for further editing. The semantic search feature allows users to input search queries and define geographic areas through map interaction. The system processes these inputs using the BGE-M3 embedding model and presents the top twenty most relevant results based on cosine similarity ranking. Users retain complete autonomy in determining which search results, if any, warrant further investigation or incorporation into their workflow.

The AI does not guide or automate any decisions — its role is strictly to assist in formatting and provide suggestions of local offerings. Only RDF outputs that pass validation against a SHACL file are displayed, preventing malformed or structurally incorrect data from being shown. While the semantic search introduces a recommendation mechanism that could potentially influence user choices through result ordering, users retain complete control over which search results to examine and whether to act upon them. The search interface presents results as suggestions rather than directives, and users can explore different geographic areas, or ignore search functionality entirely without affecting other application capabilities.

Neither AI component includes features designed to encourage overuse or create dependency. The RDF formatting tool requires explicit activation, and the semantic search operates only when users choose to engage with location-based discovery. Output from both systems is clearly distinguishable from user-generated content during the interaction phase, maintaining transparency about AI involvement in the workflow. There is no risk of manipulation or attachment behaviours developing.

Assessment: Low Risk

- It is clear which content is AI-generated
- AI-generated content is only usable in RDF format
- Search results are algorithmically ranked but presented as suggestions requiring user evaluation - Users may accept or discard AI output freely
- Both systems enhance user capabilities without replacing human judgment in resource evaluation or selection decisions
- No risk of addictive, manipulative, or persuasive AI behaviour

Recommendations:

- Consider providing a help tooltip or FAQ entry to explain how AI suggestions are validated and embedded
- Provide clear indicators that search results are algorithmically ranked based on semantic similarity and may require verification for accuracy and current availability.

2.2 Technical Robustness and Safety

The AI components have demonstrated variable performance during development phases but have not undergone systematic testing protocols. The RDF formatting component shows inconsistent output quality, producing accurate results in some cases while failing validation in others. Since only valid RDF data is accepted by the editor, invalid results are discarded before they can affect user workflows. The semantic search functionality introduces additional technical dependencies that similarly lack comprehensive testing across different query types and geographic regions.

Using the AI is completely optional. When used, the generated data appears in a preview window, and users can decide whether to embed it in the RDF graph. If the AI response is invalid or triggers an error, users are notified and may continue working without it or retry.

RDF data is stored on servers controlled by the development team using the Apache Jena framework. Access requires authentication via Keycloak. While personal data can be entered, the app is primarily designed for managing publicly available resources. The semantic search component accesses OpenStreetMap data through the Overpass API using established Leaflet integration, creating dependencies on external data quality and availability that vary by geographic location. Search result quality depends significantly on underlying OpenStreetMap data completeness and accuracy, which fluctuates across different regions and resource types. The embedding processing operates locally, reducing external service dependencies for the core semantic matching functionality, though the system still relies on Overpass API availability for geographic data retrieval.

The project will be supported by the development team for two to three years, with most updates front-loaded in the initial months. Ongoing support is minimal but planned.

Assessment: Low Risk

- AI has not been systematically tested
- No formal robustness or failover mechanisms in place, but all features are usable without the AI component.
- AI errors are handled gracefully
- Straightforward infrastructure and internal hosting mitigate many safety concerns

Recommendations:

- Conduct informal testing sessions to identify common failure cases in AI output
- Track the frequency and nature of validation failures and quality of semantic search result rankings

2.3 Privacy and Data Governance

Users are free to input any RDF-compatible data, which is a weak constraint. The system does not attempt to identify or prevent the entry of personal data, nor does it enforce any privacy policies in the data itself. The application is mainly intended for storing public resource data, but this is neither enforced nor enforceable.

The number of people with access to the shared RDF store is unclear, though likely in the tens or low hundreds. All authenticated users have equal access, as there are no role-based permission systems. This may pose risks around accidental data modification or deletion.

No mechanisms exist for data classification, consent management, or user-specific visibility settings. While the Ollama server and OpenStreetMap API usage are constrained, the system does not currently analyse input for misuse or sensitive content.

Assessment: Low Risk

- No explicit controls over personal data input or sharing
- No role-based access or audit trail for data changes
- Small and limited user base currently minimizes the impact

Recommendations:

- Consider role-based permissions
- Consider an edit history feature
- Clarify acceptable use and data input boundaries for users

2.4 Transparency

The graph editor simplifies RDF structures into a format resembling a rich table-like form. Users do not need RDF knowledge to interact with it. AI-generated content is clearly visible during the preview phase, but once embedded, it is not marked as AI-generated. Editing changes are reversible until committed.

Error messages for AI failures are minimal and may not always explain the problem clearly.

The semantic search component introduces algorithmic transparency considerations that affect user understanding of system operations. Search results undergo ranking through BGE-M3 embedding analysis and cosine similarity calculations, presenting users with algorithmically determined relevance scores. However, the interface does not explain why specific results achieve higher rankings than others, potentially creating confusion about the basis for result ordering.

Users may not recognize that search effectiveness varies significantly depending on the quality and completeness of underlying OpenStreetMap data, which differs substantially across geographic regions and resource categories.

Search results lack explicit indicators that their presentation order derives from algorithmic processing rather than chronological, alphabetical, or other conventional sorting methods.

Assessment: Low Risk

- AI-generated content not labelled after embedding
- UI is mostly transparent, but complex structures may still be confusing to some users
- Search result presentation lacks algorithmic transparency indicators, but the straightforward interface design minimises confusion

Recommendations:

- Improve validation error messages or provide examples of acceptable input
- Consider implementing interface indicators that communicate the algorithmic nature of search result ordering and provide basic explanations of semantic similarity ranking
- Consider adding contextual help information about OpenStreetMap data limitations and geographic coverage variations

2.5 Diversity, Non-discrimination and Fairness

Although the app supports work with underserved populations, those individuals are not the users themselves. As such, the UI and feature set were not explicitly designed with their direct accessibility in mind.

There is no specific review or mitigation process for potential bias in AI-generated content. Multilingual support is assumed based on the underlying model (Phi-4), but the effectiveness of different languages is yet untested. Users provide and maintain resource data independently, which introduces variability and potential underrepresentation of certain communities.

The semantic search functionality introduces potential bias mechanisms through its reliance on machine learning models and crowdsourced data sources. The BGE-M3 embedding model may demonstrate varying performance levels across different languages, cultural contexts, and types of resource descriptions. Search algorithms may systematically favour establishments or services that align more closely with the model's training data patterns, potentially underrepresenting resources that serve specific communities or utilise non-standard terminology. This algorithmic bias could disproportionately affect the discoverability of services tailored to the three target populations, particularly those that employ culturally specific language or operate within specialised community networks.

The system's dependence on OpenStreetMap data compounds potential fairness concerns through the geographic and demographic biases inherent in crowdsourced mapping platforms. Resource availability and data quality within OpenStreetMap varies significantly across different neighbourhoods, regions, and community types. Areas with lower socioeconomic status or marginalized communities often exhibit reduced mapping coverage and detail, which could result in systematic underrepresentation of relevant resources in search results. This data limitation particularly affects the discovery of community-specific services that may not receive adequate documentation in mainstream mapping platforms.

Assessment: Medium Risk

- The application's current user base and research context limit immediate discrimination risks, though systematic bias testing remains absent
- Users must explicitly approve AI-generated suggestions from both components, preserving human oversight in content decisions
- Multilingual capability is unverified

Recommendations:

- Conduct performance evaluation of both AI components across multiple EU languages, with particular attention to effectiveness for resources serving the three target populations
- Include prompts or suggestions to review AI output for fairness before accepting it

2.6 Societal and Environmental Well-being

The app is meant to positively impact link workers' ability to connect vulnerable individuals to appropriate resources. It provides a lightweight way to organize and share data, making link work more efficient. The addition of semantic search capabilities further amplifies this potential by reducing the time and effort required for resource discovery, allowing link workers to identify relevant services more quickly and comprehensively within specific geographic areas.

Risks include the potential for incorrect or low-quality information to spread within the user base. There is no validation for correctness beyond the RDF format, and no version control or moderation features currently exist. The semantic search component may inadvertently promote certain types of

resources over others through its ranking mechanisms, potentially affecting the diversity of connections made by link workers. Results depend on the completeness and accuracy of underlying OpenStreetMap data, which varies significantly across different geographic regions and service categories.

Environmentally, the system is considered low impact. The Ollama AI server is hosted locally and queried on demand. No large-scale infrastructure or frequent inference operations are involved.

Assessment: Low Risk

- Designed to benefit link workers and vulnerable groups indirectly
- Lightweight infrastructure, minimal energy footprint
- Some risk of low-quality data propagation

Recommendations:

- Add optional moderation or versioning features
- Document best practices for verifying resource information
- Maintain lightweight deployment even if the system scales

2.7 Accountability

The development team currently handles all technical responsibilities, including hosting, bug fixes, and maintenance. However, responsibilities across the broader project consortium are not clearly defined, which may lead to ambiguity if governance issues arise.

Users can report issues informally via Redmine or through personal contact with the team, but there is no direct feedback channel within the application. This may be sufficient for the research context but could become limiting if the app is extended.

There is no moderation or flagging system for problematic data, and no audit logs to trace AI usage or user activity.

Assessment: Medium Risk

- Clear internal responsibility, but unclear external roles
- No in-app issue reporting or user-facing contact
- No audit trails or moderation mechanisms

Recommendations:

- Define project-wide roles and responsibilities for data and system management
- Add a simple contact or feedback form within the app
- Consider lightweight user logging or change history tracking for accountability

3. Conclusion

The ALTAI assessment indicates a generally low risk profile. The system includes clear safeguards, requiring explicit user activation for all AI functions, limiting AI-generated outputs to suggestions rather than decisions, and ensuring transparent AI involvement during user interactions. Additionally, invalid data is automatically excluded, and the infrastructure is locally hosted with minimal external dependencies. While the system shows promise in enhancing social prescribing efforts for vulnerable populations, medium-risk concerns remain around fairness, multilingual effectiveness, and accountability structures. Addressing these areas through targeted improvements will strengthen the application's trustworthiness and long-term viability.