



Intelligent Automation in Anonymization for CTIS Submission

Pinku Hooda
Hiteshkumar Raval

www.geninvo.com
North America | UK | Europe | India

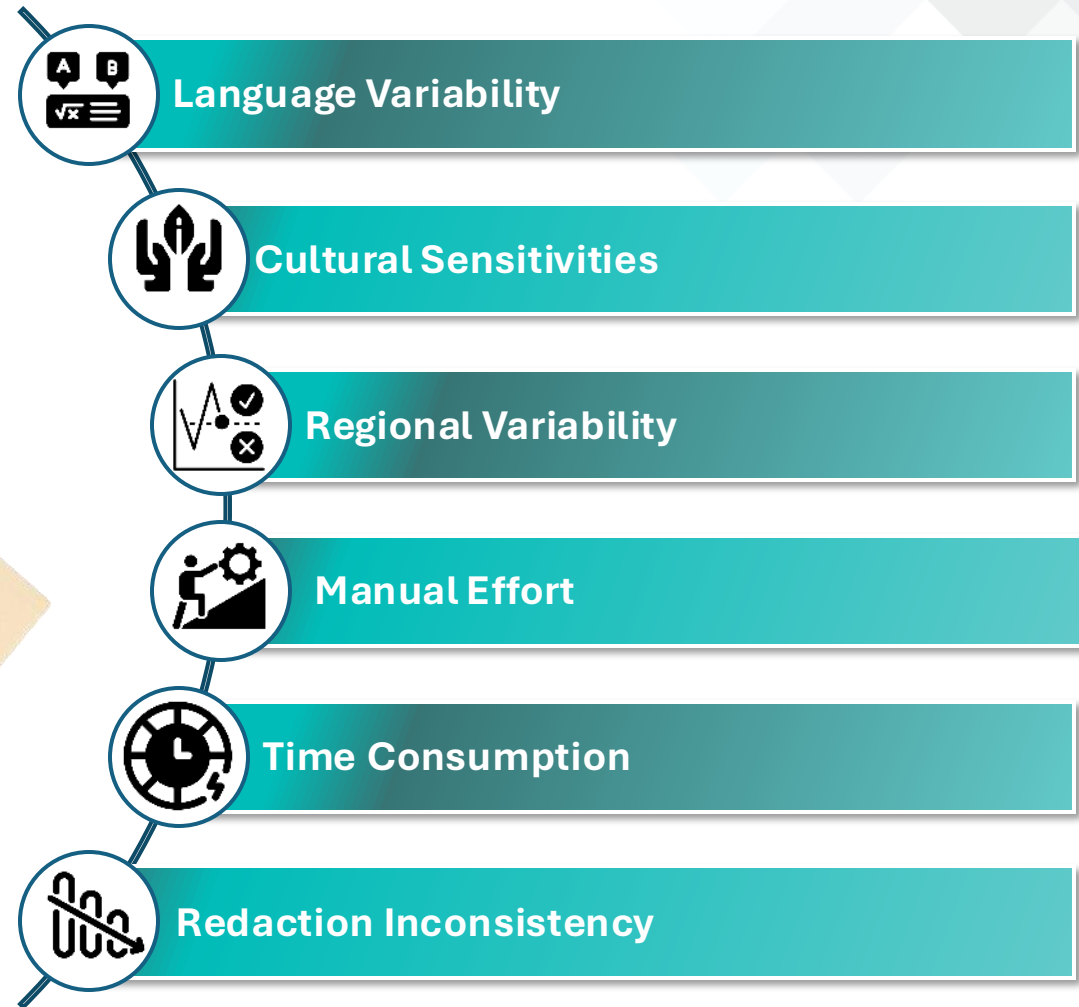


Agenda

- 1** Overview of Current Redaction Process
- 2** Challenges with the Current Redaction Process
- 3** Complexities in Automating Redaction
- 4** Technological Solutions: NLP & Machine Translation
- 5** Automation Approach For Multilingual Redaction
- 6** Examples of Multilingual Redaction
- 7** The Future of Automated Anonymization
- 8** Q&A



Challenges with the Current Redaction Process



Complexities in Automating Redaction

Redacting Using English as a Reference Point

- English redactions may not translate directly into other languages, requiring systems to identify equivalent sensitive information in translated documents.
- Ensuring that redaction practices in the English version align with equivalent meanings in non-English versions is complex.



Legal and Compliance Complexities

- Legal and regulatory redaction requirements vary across regions (e.g., GDPR in the EU vs. HIPAA in the U.S.).
- Different jurisdictions define sensitive data differently, such as race, ethnicity, health, or political opinions.
- Redaction systems must be adaptable to regional legal frameworks to ensure compliance.
- Ensuring this adaptability is a key challenge for automating the redaction process.

Technological Solutions: NLP & Machine Translation

Natural Language Processing (NLP)

- NLP enables machines to understand, interpret, and generate human language, including the identification of sensitive information.
- Automates the identification and redaction of sensitive information in documents.
- Processes context, syntax, and semantics to ensure correct redaction across languages.
- Key NLP Components:
 - Named Entity Recognition (NER)
 - Semantic Analysis
 - Multilingual NLP Models



Technological Solutions: NLP & Machine Translation

Named Entity Recognition (NER)

- Identifies and redacts sensitive entities (e.g., names, medical conditions, addresses) in documents.
- Ensures compliance with privacy regulations.
- Example: Redacting "John Smith" and "heart disease" in clinical documents.

Semantic Analysis

- Overcomes limitations of traditional keyword matching techniques.
- Avoids over-redaction and under-redaction.
- Understands the underlying meaning of text, ensuring redactions are contextually appropriate.
- Example: In clinical trial documents, terms like "heart disease" or "sensitive medical history" are flagged for redaction, ensuring sensitive data is protected.

Multilingual NLP Models

- Models like BERT process multiple languages, handling grammatical and syntactical differences.
- Essential for ensuring accurate detection and redaction of sensitive data across both original and translated documents.
- Preserves redactions in multilingual environments, ensuring data protection across languages.



Technological Solutions: NLP & Machine Translation

Machine Translation for Multilingual Redaction

- Preserves redactions across languages, ensuring the integrity of sensitive information maintained in translated documents.
- Machine translation tools like Google Translate, DeepL, and domain-specific systems help translate redacted documents into languages with varying sentence structures (e.g., Spanish).
- Ensure redactions remain intact, and context is preserved, despite the differences in syntax.

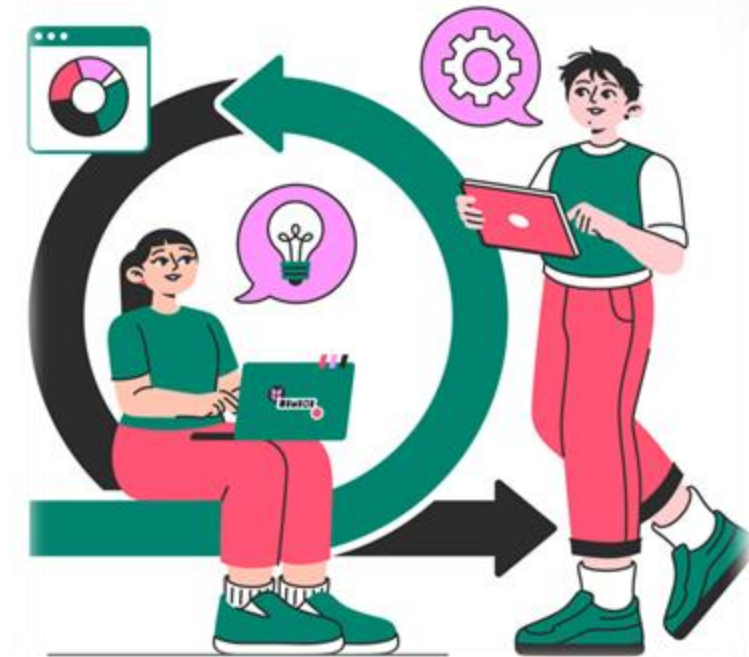
Example of Redacted Translation

- **Original English Text:**
"John Smith, a 45-year-old male with a history of heart disease, was admitted."
- **After Redaction in English:**
"[REDACTED], a 45-year-old male with a history of [REDACTED], was admitted."
- **After Machine Translation to Spanish:**
"[REDACTED], un hombre de 45 años con antecedentes de [REDACTED], fue admitido."

Technological Solutions: NLP & Machine Translation

Machine Learning Feedback Loops

- Crucial for improving redaction system accuracy and adaptability in multilingual settings.
- Continuous model improvement through feedback from stakeholders (e.g., regulatory experts, linguists).
- Handling language-specific challenges, For example, recognizing sensitive medical terms like “hipertensión” in Spanish and adapting to emerging terminology like COVID-19-related terms.
- Adaptability to changing data protection regulations and emerging industry standards.



Automation Approach For Multilingual Redaction

Translation API Process & Enhancements with AI/ML

Document Parsing and Segmentation

- **Fitz (pyMUPDF) Library** to extract and segment both the English and target language documents into blocks, each associated with specific coordinates.
- Enables structured processing of document content.
- **Feature extraction** helps represent text into a format suitable for further analysis and processing.

Preprocessing and Data Cleaning

- **Data preprocessing techniques** to clean data by removing unwanted characters, handling null values, and eliminating duplicates from redacted strings.
- Crucial in data preparation for **machine learning models** to ensure clean, noise-free data, improving analysis and prediction accuracy.

String Processing

- **re(regular expression) package**, for **text normalization** and **string manipulation**, addressing punctuation, newline characters, and sentence boundaries.
- Standardizing text into a consistent format for further analysis & ingestion into machine learning models.
- Essential for **natural language processing (NLP)** tasks such as **tokenization** and **vectorization**.

Automation Approach For Multilingual Redaction

Translation API Process & Enhancements with AI/ML

Translation and NLP

- **Amazon Translate API** for **automated language translation**, powered by AI, to translate the redacted text from English to the target language.
- Translation process further enhanced by **machine translation models** that incorporate statistical learning and deep learning algorithms.

Text Embedding & Semantic Similarity Calculation

- **Sentence-Transformer model (intfloat/multilingual-e5-base)** that generates **sentence embeddings**, to compare the meaning and context of sentences in different languages.
- Embeddings represent sentences in a **vector space**, enabling **semantic similarity analysis** through **cosine similarity** and **Jaccard similarity metrics**.

```
from sentence_transformers import SentenceTransformer, util  
bert_model = SentenceTransformer('intfloat/multilingual-e5-base')
```

- Applying **vector space models** and calculating **similarity scores** to determine relevant matches between translated sentences and their counterparts in target language documents.
- Leveraging **machine learning techniques** for prediction and pattern recognition.

Automation Approach For Multilingual Redaction

Translation API Process & Enhancements with AI/ML

Semantic Analysis using WordNet

- Integrate **WordNet** from **NLTK** to perform **semantic analysis**.
- Leveraging AI-driven **word embeddings** & **synonym matching** to enhance model's understanding of word relationships and enabling more accurate matching of redacted strings even with synonyms or context-based variations.
- Apply AI to identify these semantic connections between words, improving overall translation quality and consistency.

```
from nltk.corpus import wordnet as wn
```

- Additionally, **Open Multilingual Wordnet (OMW)** to enhance model's ability to handle multiple languages, supporting **cross-lingual word relations** for improved multilingual processing.

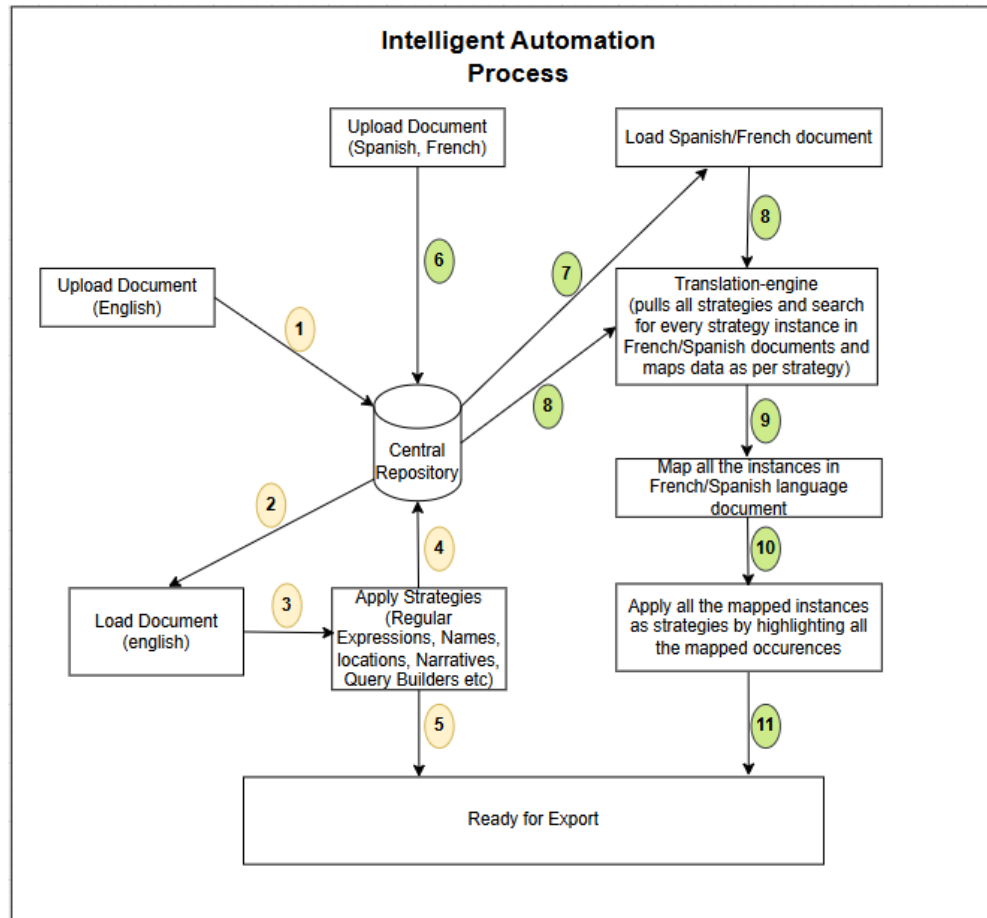
```
import nltk  
nltk.download('omw-1.4')
```

Intersection Identification Using AI Models

- **AI-based techniques** to identify the intersection between the target string/word and the corresponding sentence using **similarity scores** generated from **sentence embeddings** and **TF-IDF vectors**.
- Utilize **machine learning** to determine the most relevant matches by calculating the semantic distance between sentences, ensuring precise redacted string alignment across documents.

Automation Approach For Multilingual Redaction

Intelligent Automation Process Flow



Step 1 – Step 5:

- The **English document** is uploaded to the central repository.
- It is processed by applying **redaction strategies** to generate the redacted English document.
- Any **modifications** to the redaction strategies at the English document level are stored back in the central repository.

Step 6 – Step 11:

- The **Other Language (French/Spanish)** document is uploaded to the central repository.
- The **translation engine** pulls all redaction strategies applied to the English document and processes the other language document by applying these strategies.
- The **redacted document** in the other language is generated and all **mapped occurrences** are highlighted.



Examples of Multilingual Redaction

Redaction in English Document

1. SYNOPSIS

Name of Study Drug PA28350
Name of Active Ingredient Recombinant human monoclonal antibody (immunoglobulin gamma-1 with kappa light chains, IgG1 kappa) directed against Tumor necrosis factor-like Ligand 1A (TL1A)
Title of Study A Phase 2, Multicenter, Double-blind, Two-arm Study of Subcutaneous PA28350 for the Treatment of Subjects with Moderate to Severe Active Crohn's Disease
Study Center(s) This study will be conducted in up to approximately 150 sites worldwide. Worldwide Clinical Trials, a contract research organization, will oversee operational aspects of this study on behalf of the Sponsor of the study.
Phase of Development Phase 2

(Preview Copy)

1. SYNOPSIS

Name of pPD PA28350
Name of Active Ingredient Recombinant human monoclonal antibody (immunoglobulin gamma-1 with kappa light chains, IgG1 kappa) directed against Tumor necrosis factor-like Ligand 1A (TL1A)
Title of pPD A Phase 2, Multicenter, Double-blind, Two-arm pPD of Subcutaneous PA28350 for the Treatment of Subjects with Moderate to Severe Active Crohn's Disease
pPD Center(s) This pPD will be conducted in up to approximately 150 sites worldwide. Worldwide Clinical Trials, a contract research organization, will oversee operational aspects of this pPD on behalf of the Sponsor of the pPD.
Phase of Development Phase 2

(Redacted Copy)

*** English document where the text values marked for redaction are the words “Study” and “Drug” ***



GENINVO™

www.geninvo.com
North America | UK | Europe | India

Examples of Multilingual Redaction

Redaction in Spanish Document

1.	SINOPSIS
Nombre del fármaco del estudio PA28350	
Nombre del principio activo Anticuerpo monoclonal humano recombinante (inmunoglobulina gamma-I con cadenas ligeras kappa, IgG1 kappa) dirigido contra el ligando 1A (TL1A) similar al factor de necrosis tumoral	
Título del estudio Estudio de fase II, multicéntrico, doble ciego y de dos grupos sobre PA28350 subcutáneo para el tratamiento de sujetos con enfermedad de Crohn activa moderada a grave	
Centro(s) del estudio Este estudio se llevará a cabo en aproximadamente un máximo de 150 centros de todo el mundo. Worldwide Clinical Trials, una organización de investigación por contrato, supervisará los aspectos operativos de este estudio en nombre de el promotor del estudio.	
Fase de desarrollo Fase II	

(Preview Copy)

1.	SINOPSIS
Nombre del PPD del PPD PA28350	
Nombre del principio activo Anticuerpo monoclonal humano recombinante (inmunoglobulina gamma-I con cadenas ligeras kappa, IgG1 kappa) dirigido contra el ligando 1A (TL1A) similar al factor de necrosis tumoral	
Título del PPD PPD de fase II, multicéntrico, doble ciego y de dos grupos sobre PA28350 subcutáneo para el tratamiento de sujetos con enfermedad de Crohn activa moderada a grave	
Centro(s) del PPD Este PPD se llevará a cabo en aproximadamente un máximo de 150 centros de todo el mundo. Worldwide Clinical Trials, una organización de investigación por contrato, supervisará los aspectos operativos de este PPD en nombre de el promotor del PPD	
Fase de desarrollo Fase II	

(Redacted Copy)

*** Spanish document where redaction strategies are imported from the English version of the document ***

The Future of Automated Anonymization



Real-Time & Adaptive Anonymization

**Integration with Regulatory
Submission Workflows**

**Differential Privacy & Federated
Learning for Secure Data Sharing**

Synthetic Data & Privacy-Preserving AI

**Automated Regulatory Compliance
Across Global Jurisdictions**

Q&A





Ideas
Delivered
Intuitively

Thank You!

+1-833-GEN-INVO (436-4686) | info@geninvo.com | www.geninvo.com

