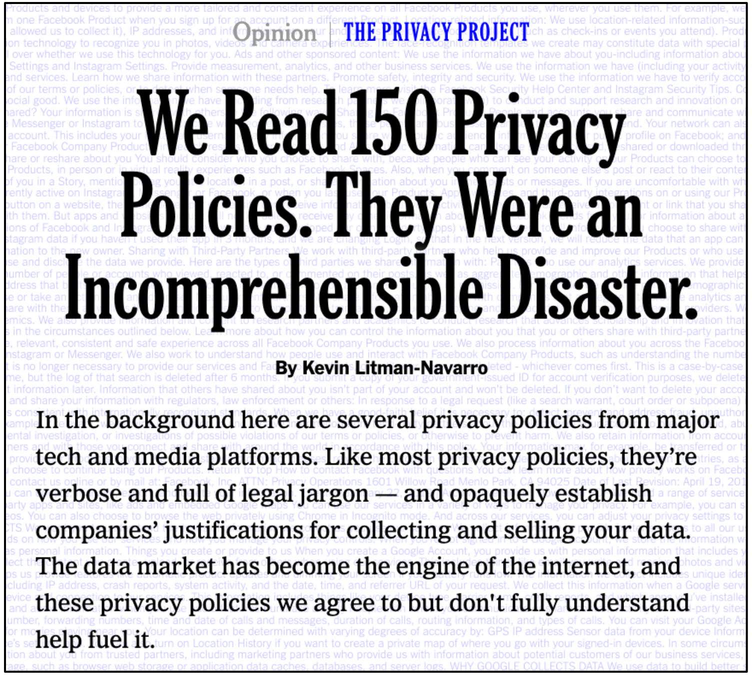
A Human-in-the-Loop Approach for Information Extraction from Privacy Policies under Data Scarcity

Michael Gebauer, Faraz Maschhur, Nicola Leschke, Elias Grünewald, and Frank Pallas

Information Systems Engineering
TU Berlin



Opinion | THE PRIVACY PROJECT

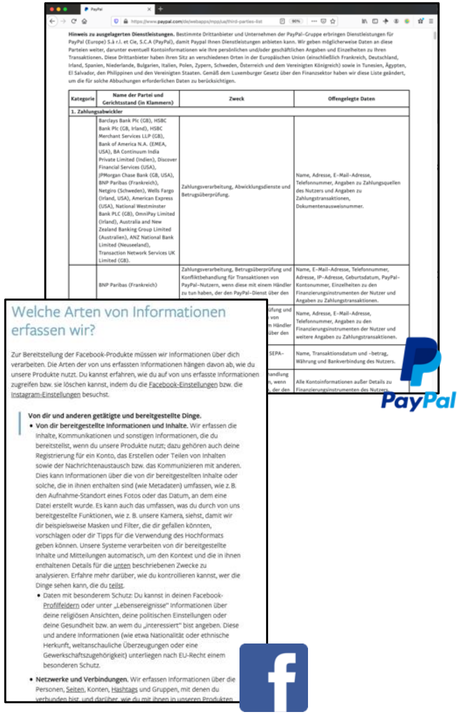
We Read 150 Privacy Policies. They Were an Incomprehensible Disaster.

By Kevin Litman-Navarro

In the background here are several privacy policies from major tech and media platforms. Like most privacy policies, they're verbose and full of legal jargon — and opaquely establish companies' justifications for collecting and selling your data. The data market has become the engine of the internet, and these privacy policies we agree to but don't fully understand help fuel it.

https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html

https://www.designboom.com/readers/dima-yarovinsky-visualizes-facebook-instagram-snapchat-terms-of-service-05-07-2018/



https://www.paypal.com/de/webapps/mpp/ua/third-parties-list

Hinweis zu ausgelagerten Dienstleistungen. Bestimmte Drittanbieter und Unternehmen der PayPal-Gruppe erbringen Dienstleistungen für PayPal (Europe) S.à r.l. et Cie, S.C.A (PayPal), damit Paypal Ihnen Dienstleistungen anbieten kann. Wir geben möglicherweise Daten an diese Parteien weiter, darunter eventuell Kontoinformationen wie Ihre persönlichen und/oder geschäftlichen Angaben und Einzelheiten zu Ihren Transaktionen. Diese Drittanbieter haben ihren Sitz an verschiedenen Orten in der Europäischen Union (einschließlich Frankreich, Deutschland, Irland, Spanien, Niederlande, Bulgarien, Italien, Polen, Zypern, Schweden, Österreich und dem Vereinigten Königreich) sowie in Tunesien, Ägypten, El Salvador, den Philippinen und den Vereinigten Staaten. Gemäß dem Luxemburger Gesetz über den Finanzsektor haben wir diese Liste geändert, um die für solche Abbuchungen erforderlichen Daten zu berücksichtigen.

Kategorie	Name der Partei und Gerichtsstand (in Klammern)	Zweck	Offengelegte Daten
1. Zahlungsabwickler			
	Barclays Bank Plc (GB), HSBC Bank Plc (GB, Irland), HSBC Merchant Services LLP (GB), Bank of America N.A. (EMEA, USA), BA Continuum India Private Limited (Indien), Discover Financial Services (USA), JPMorgan Chase Bank (GB, USA), BNP Paribas (Frankreich), Netgiro (Schweden), Wells Fargo (Irland, USA), American Express (USA), National Westminster Bank PLC (GB), OmniPay Limited (Irland), Australia and New Zealand Banking Group Limited (Australien), ANZ National Bank Limited (Neuseeland), Transaction Network Services UK Limited (GB).	Zahlungsverarbeitung, Abwicklungsdienste und Betrugsüberprüfung.	Name, Adresse, E-Mail-Adresse, Telefonnummer, Angaben zu Zahlungsquellen des Nutzers und Angaben zu Zahlungstransaktionen, Dokumentenausweisnummer.
	BNP Paribas (Frankreich)	Zahlungsverarbeitung, Betrugsüberprüfung und Konfliktbehandlung für Transaktionen von PayPal-Nutzern, wenn diese mit einem Händler zu tun haben, der den PayPal-Dienst über den	Name, E-Mail-Adresse, Telefonnummer, Adresse, IP-Adresse, Geburtsdatum, PayPal-Kontonummer, Einzelheiten zu den Finanzierungsinstrumenten der Nutzer und Angaben zu Zahlungstransaktionen.
		üfung und von m Händler über den	Name, Adresse, E-Mail-Adresse, Telefonnummer, Angaben zu den Finanzierungsinstrumenten der Nutzer und weitere Angaben zu Zahlungstransaktionen.
		SEPA-	Name, Transaktionsdatum und -betrag, Währung und Bankverbindung des Nutzers.
		handlung n, wenn , der den	Alle Kontoinformationen außer Details zu Finanzierungsinstrumenten des Nutzers.

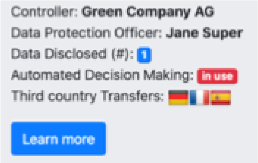
Welche Arten von Informationen erfassen wir?

Zur Bereitstellung der Facebook-Produkte müssen wir Informationen über dich verarbeiten. Die Arten der von uns erfassten Informationen hängen davon ab, wie du unsere Produkte nutzt. Du kannst erfahren, wie du auf von uns erfasste Informationen zugreifen bzw. sie löschen kannst, indem du die Facebook-Einstellungen bzw. die Instagram-Einstellungen besuchst.

Von dir und anderen getätigte und bereitgestellte Dinge.

- **Von dir bereitgestellte Informationen und Inhalte.** Wir erfassen die Inhalte, Kommunikationen und sonstigen Informationen, die du bereitstellst, wenn du unsere Produkte nutzt; dazu gehören auch deine Registrierung für ein Konto, das Erstellen oder Teilen von Inhalten sowie der Nachrichtenaustausch bzw. das Kommunizieren mit anderen. Dies kann Informationen über die von dir bereitgestellten Inhalte oder solche, die in ihnen enthalten sind (wie Metadaten) umfassen, wie z. B. den Aufnahme-Standort eines Fotos oder das Datum, an dem eine Datei erstellt wurde. Es kann auch das umfassen, was du durch von uns bereitgestellte Funktionen, wie z. B. unsere Kamera, siehst, damit wir dir beispielsweise Masken und Filter, die dir gefallen könnten, vorschlagen oder dir Tipps für die Verwendung des Hochformats geben können. Unsere Systeme verarbeiten von dir bereitgestellte Inhalte und Mitteilungen automatisch, um den Kontext und die in ihnen enthaltenen Details für die unten beschriebenen Zwecke zu analysieren. Erfahre mehr darüber, wie du kontrollieren kannst, wer die Dinge sehen kann, die du teilst.
  - Daten mit besonderem Schutz: Du kannst in deinen Facebook-Profilfeldern oder unter „Lebensereignisse" Informationen über deine religiösen Ansichten, deine politischen Einstellungen oder deine Gesundheit bzw. an wem du „interessiert" bist angeben. Diese und andere Informationen (wie etwa Nationalität oder ethnische Herkunft, weltanschauliche Überzeugungen oder eine Gewerkschaftszugehörigkeit) unterliegen nach EU-Recht einem besonderen Schutz.
- **Netzwerke und Verbindungen.** Wir erfassen Informationen über die Personen, Seiten, Konten, Hashtags und Gruppen, mit denen du verbunden bist, und darüber, wie du mit ihnen in unseren Produkten

?



Controller: **Green Company AG**

Data Protection Officer: **Jane Super**

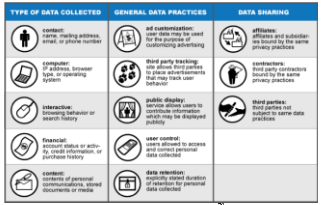
Data Disclosed (#): 1

Automated Decision Making: in use

Third country Transfers: 🇩🇪 🇫🇷 🇪🇸

Browser-Plugins

Visualising Data-Sharing-Networks

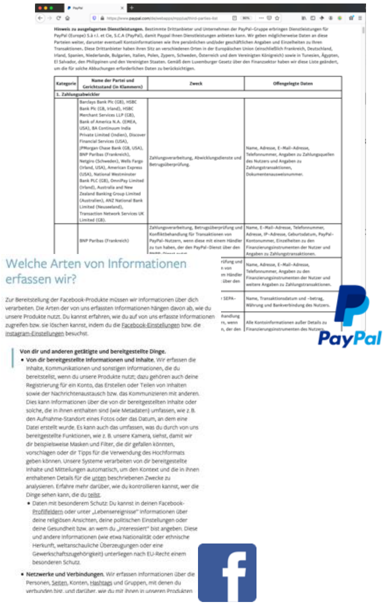


TYPE OF DATA COLLECTED	GENERAL DATA PRACTICES	DATA SHARING
contact: name, mailing address, email, or phone number	**ad customization:** user data may be used for the purpose of customizing advertising	**affiliates:** affiliates and subsidiaries bound by the same privacy practices
computer: IP address, browser type, or operating system	**third party tracking:** site allows third parties to place advertisements that may track user behavior	**contractors:** third party contractors bound by the same privacy practices
interactive: browsing behavior or search history	**public display:** service allows users to contribute information which may be displayed publicly	**third parties:** third parties not subject to same data practices
financial: account status or activity, credit information, or purchase history	**user control:** users allowed to access and correct personal data collected	
content: contents of personal communications, stored documents or media	**data retention:** explicitly stated duration of retention for personal data collected	

Privacy Icons

Grünewald et al. (2023). „Enabling Versatile Privacy Interfaces Using Machine-Readable Transparency Information“. In Privacy Symposium 2023

Indispensable: machine-readable representation



Hinweis zu ausgelagerten Dienstleistungen. Bestimmte Drittanbieter und Unternehmen der PayPal-Gruppe erbringen Dienstleistungen für PayPal (Europe) S.à r.l. et Cie, S.C.A (PayPal), damit Paypal Ihnen Dienstleistungen anbieten kann. Wir geben möglicherweise Daten an diese Parteien weiter, darunter eventuell Kontoinformationen wie Ihre persönlichen und/oder geschäftlichen Angaben und Einzelheiten zu Ihren Transaktionen. Diese Drittanbieter haben ihren Sitz an verschiedenen Orten in der Europäischen Union (einschließlich Frankreich, Deutschland, Irland, Spanien, Niederlande, Bulgarien, Italien, Polen, Zypern, Schweden, Österreich und dem Vereinigten Königreich) sowie in Tunesien, Ägypten, El Salvador, den Philippinen und den Vereinigten Staaten. Gemäß dem Luxemburger Gesetz über den Finanzsektor haben wir diese Liste geändert, um die für solche Abbuchungen erforderlichen Daten zu berücksichtigen.

Kategorie	Name der Partei und Gerichtsstand (in Klammern)	Zweck	Offengelegte Daten
1. Zahlungsabwickler			
	Barclays Bank Plc (GB), HSBC Bank Plc (GB, Irland), HSBC Merchant Services LLP (GB), Bank of America N.A. (EMEA, USA), BA Continuum India Private Limited (Indien), Discover Financial Services (USA), JPMorgan Chase Bank (GB, USA), BNP Paribas (Frankreich), Netgiro (Schweden), Wells Fargo (Irland, USA), American Express (USA), National Westminster Bank PLC (GB), OmniPay Limited (Irland), Australia and New Zealand Banking Group Limited (Australien), ANZ National Bank Limited (Neuseeland), Transaction Network Services UK Limited (GB).	Zahlungsverarbeitung, Abwicklungsdienste und Betrugsüberprüfung.	Name, Adresse, E-Mail-Adresse, Telefonnummer, Angaben zu Zahlungsquellen des Nutzers und Angaben zu Zahlungstransaktionen, Dokumentenausweisnummer.
	BNP Paribas (Frankreich)	Zahlungsverarbeitung, Betrugsüberprüfung und Konfliktbehandlung für Transaktionen von PayPal-Nutzern, wenn diese mit einem Händler zu tun haben, der den PayPal-Dienst über den BNPP-Dienst nutzt.	Name, E-Mail-Adresse, Telefonnummer, Adresse, IP-Adresse, Geburtsdatum, PayPal-Kontonummer, Einzelheiten zu den Finanzierungsinstrumenten der Nutzer und Angaben zu Zahlungstransaktionen.
		…rüfung und …n von …m Händler … über den	Name, Adresse, E-Mail-Adresse, Telefonnummer, Angaben zu den Finanzierungsinstrumenten der Nutzer und weitere Angaben zu Zahlungstransaktionen.
		… SEPA-	Name, Transaktionsdatum und -betrag, Währung und Bankverbindung des Nutzers.
		…handlung …n, wenn …n, der den	Alle Kontoinformationen außer Details zu Finanzierungsinstrumenten des Nutzers.

Welche Arten von Informationen erfassen wir?

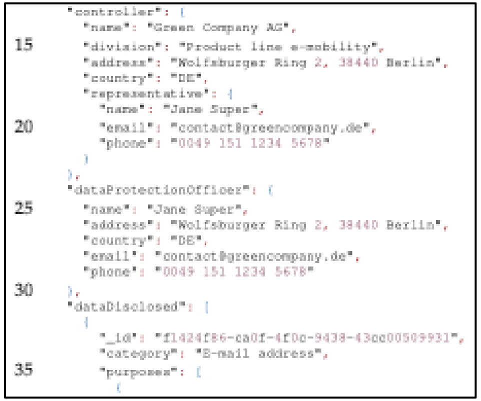
Zur Bereitstellung der Facebook-Produkte müssen wir Informationen über dich verarbeiten. Die Arten der von uns erfassten Informationen hängen davon ab, wie du unsere Produkte nutzt. Du kannst erfahren, wie du auf von uns erfasste Informationen zugreifen bzw. sie löschen kannst, indem du die Facebook-Einstellungen bzw. die Instagram-Einstellungen besuchst.

Von dir und anderen getätigte und bereitgestellte Dinge.

- **Von dir bereitgestellte Informationen und Inhalte.** Wir erfassen die Inhalte, Kommunikationen und sonstigen Informationen, die du bereitstellst, wenn du unsere Produkte nutzt; dazu gehören auch deine Registrierung für ein Konto, das Erstellen oder Teilen von Inhalten sowie der Nachrichtenaustausch bzw. das Kommunizieren mit anderen. Dies kann Informationen über die von dir bereitgestellten Inhalte oder solche, die in ihnen enthalten sind (wie Metadaten) umfassen, wie z. B. den Aufnahme-Standort eines Fotos oder das Datum, an dem eine Datei erstellt wurde. Es kann auch das umfassen, was du durch von uns bereitgestellte Funktionen, wie z. B. unsere Kamera, siehst, damit wir dir beispielsweise Masken und Filter, die dir gefallen könnten, vorschlagen oder dir Tipps für die Verwendung des Hochformats geben können. Unsere Systeme verarbeiten von dir bereitgestellte Inhalte und Mitteilungen automatisch, um den Kontext und die in ihnen enthaltenen Details für die unten beschriebenen Zwecke zu analysieren. Erfahre mehr darüber, wie du kontrollieren kannst, wer die Dinge sehen kann, die du teilst.
  - Daten mit besonderem Schutz: Du kannst in deinen Facebook-Profilfeldern oder unter „Lebensereignisse" Informationen über deine religiösen Ansichten, deine politischen Einstellungen oder deine Gesundheit bzw. an wem du „Interessiert" bist angeben. Diese und andere Informationen (wie etwa Nationalität oder ethnische Herkunft, weltanschauliche Überzeugungen oder eine Gewerkschaftszugehörigkeit) unterliegen nach EU-Recht einem besonderen Schutz.
- **Netzwerke und Verbindungen.** Wir erfassen Informationen über die Personen, Seiten, Konten, Hashtags und Gruppen, mit denen du verbunden bist, und darüber, wie du mit ihnen in unseren Produkten

⟷

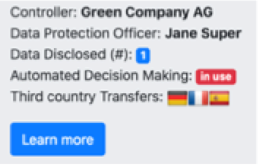
Transparency Information Language & Toolkit



```
    "controller": {
      "name": "Green Company AG",
      "division": "Product line e-mobility",
      "address": "Wolfsburger Ring 2, 38440 Berlin",
      "country": "DE",
      "representative": {
        "name": "Jane Super",
        "email": "contact@greencompany.de",
        "phone": "0049 151 1234 5678"
      }
    },
    "dataProtectionOfficer": {
      "name": "Jane Super",
      "address": "Wolfsburger Ring 2, 38440 Berlin",
      "country": "DE",
      "email": "contact@greencompany.de",
      "phone": "0049 151 1234 5678"
    },
    "dataDisclosed": [
      {
        "_id": "f1424f86-ca0f-4f0c-9438-43cc00509931",
        "category": "E-mail address",
        "purposes": [
          {
```

Grünewald and Pallas (2021). „TILT: A GDPR-Aligned Transparency Information Language and Toolkit for Practical Privacy Engineering ". In ACM FAccT 2021.

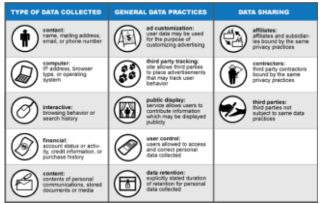
⟷



Controller: **Green Company AG**
Data Protection Officer: **Jane Super**
Data Disclosed (#): 1
Automated Decision Making: in use
Third country Transfers: 🇩🇪 🇫🇷 🇪🇸

Browser-Plugins

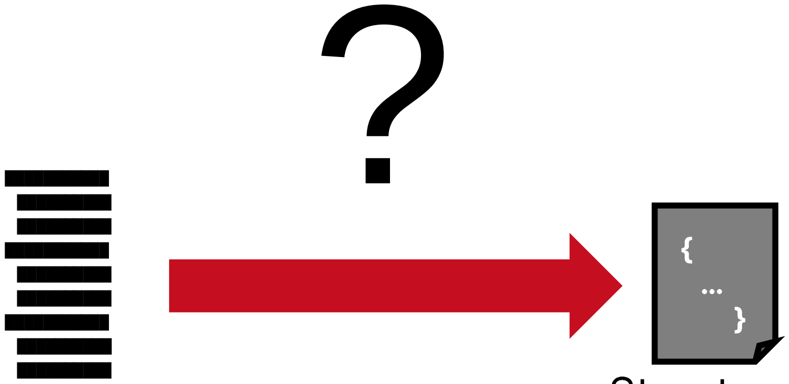
Visualising Data-Sharing-Networks



TYPE OF DATA COLLECTED	GENERAL DATA PRACTICES	DATA SHARING
contact: name, mailing address, email, or phone number	**ad customization:** user data may be used for the purpose of customizing advertising	**affiliates:** affiliates and subsidiaries bound by the same privacy practices
computer: IP address, browser type, or operating system	**third party tracking:** site allows third parties to place advertisements that may track user behavior	**contractors:** third party contractors bound by the same privacy practices
interactive: browsing behavior or search history	**public display:** service allows users to contribute information which may be displayed publicly	**third parties:** third parties not subject to same data practices
financial: account status or activity, credit information, or purchase history	**user control:** users allowed to access and correct personal data collected	
content: contents of personal communications, stored documents or media	**data retention:** explicitly stated duration of retention for personal data collected	

Privacy Icons

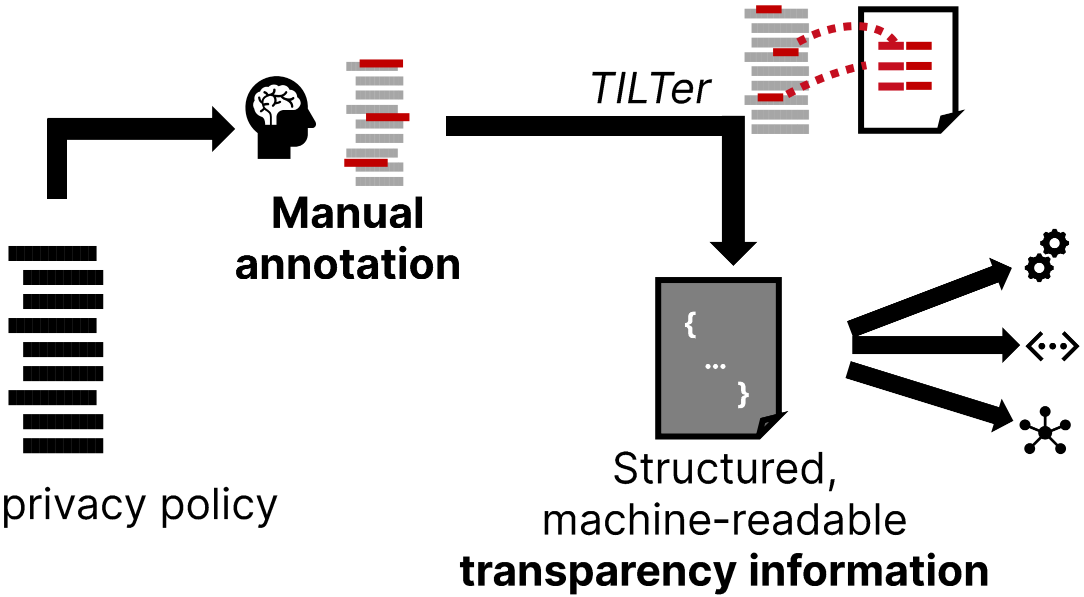
How to get TILT?



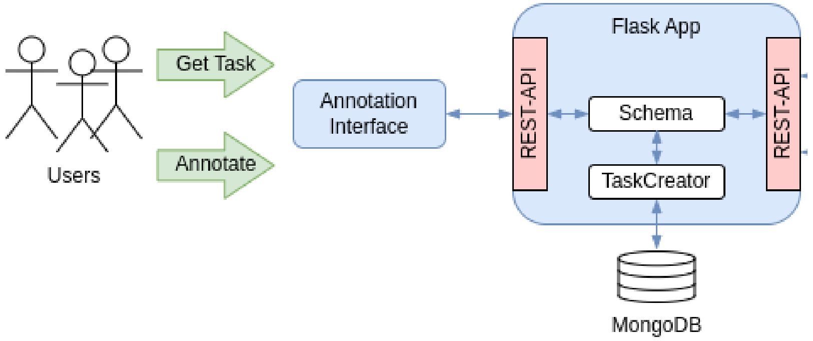
textual privacy policy

Structured,
machine-readable
transparency information

General Approach



TILTer


Users
Get Task
Annotate
Annotation Interface
Flask App
REST-API
Schema
TaskCreator
REST-API
MongoDB

Annotation Interface

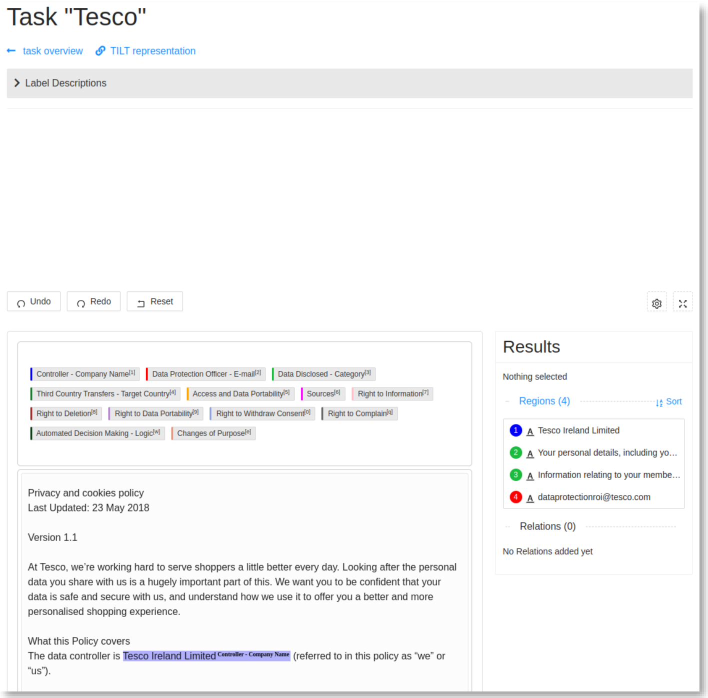
Hierarchically structured **labels** for:

- Tokens, e.g.
  - DPO (name, e-mail, address, ...)
  - categories
  - purpose
- Sentences, e.g.
  - data subject rights
  - automated decision making
  - changes

Multiple **tasks** per policy:

- Identify labels for a specific hierarchical level

→ still, very tedious work!



Task "Tesco"
task overview
TILT representation
Label Descriptions
Undo
Redo
Reset
Controller - Company Name[1]
Data Protection Officer - E-mail[2]
Data Disclosed - Category[3]
Third Country Transfers - Target Country[4]
Access and Data Portability[5]
Sources[6]
Right to Information[7]
Right to Deletion[8]
Right to Data Portability[9]
Right to Withdraw Consent[0]
Right to Complain[q]
Automated Decision Making - Logic[w]
Changes of Purpose[e]
Results
Regions (4)
Sort
Relations (0)

Privacy and cookies policy
Last Updated: 23 May 2018

Version 1.1

At Tesco, we're working hard to serve shoppers a little better every day. Looking after the personal data you share with us is a hugely important part of this. We want you to be confident that your data is safe and secure with us, and understand how we use it to offer you a better and more personalised shopping experience.

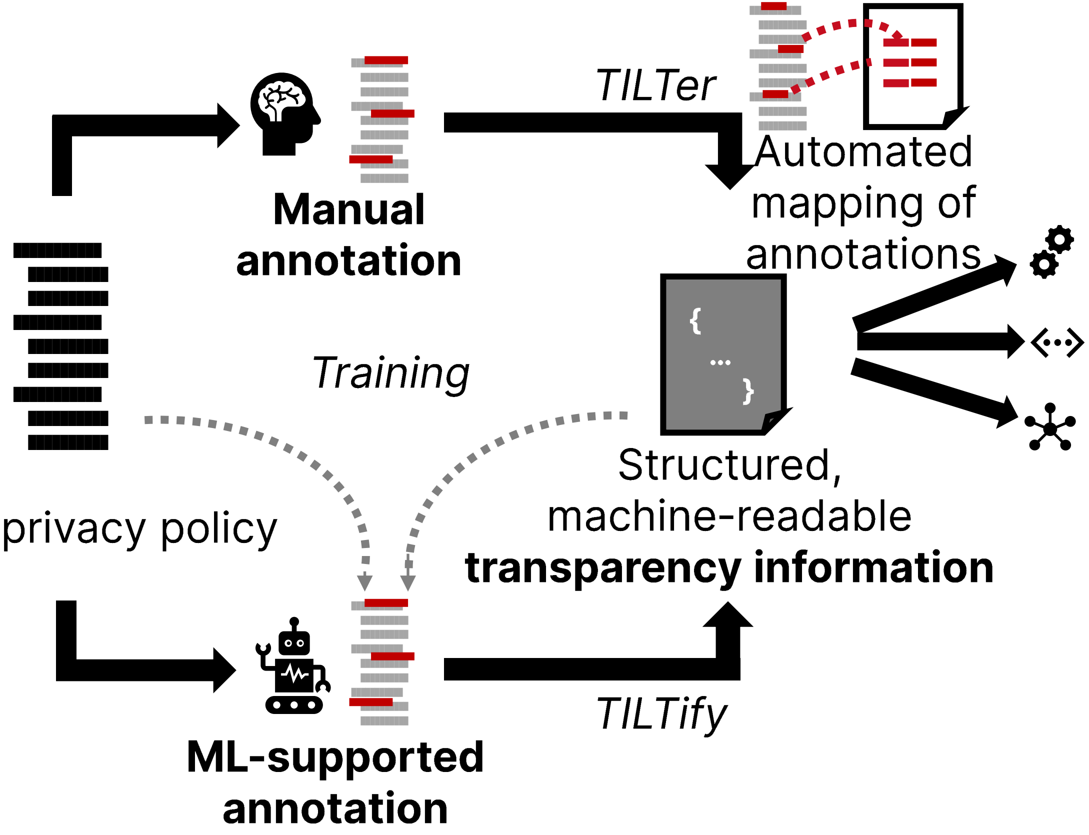
What this Policy covers
The data controller is Tesco Ireland Limited Controller - Company Name (referred to in this policy as "we" or "us").

Nothing selected

1. Tesco Ireland Limited
2. Your personal details, including yo...
3. Information relating to your membe...
4. dataprotectionroi@tesco.com

No Relations added yet

NLP Coming to the Rescue!



ML-supported Annotation of Privacy Policies

OPP-115 data set (Wilson et al. 2016)

- Polisis (Harkous et al. 2018)
- TLDR (Alabduljabbar et al. 2021)

✓ reliable classification (F1: 0.83/0.91)
✓ improving navigability

× data scarcity
× creating detailed and accurate machine-readable representations

Refining the Classification Task

Annotations for:

- Tokens, e.g.
  - DPO
  - categories
  - purpose
- Sentences, e.g.
  - data subject rights
  - automated decision making
  - changes

“We therefore would like to inform you about data privacy for websites under *.tu.berlin.”

✓

“You can obtain information about the data we have stored about you free of charge at any time without having to give a reason.”

⚡

“You also have the right to object at any time to the processing of personal data concerning you that is carried out on the basis of Article 6 paragraph 1 lit. e and f EU-GDPR.”

https://www.tu.berlin/en/data-protection

Sentence-based Information Retrieval

set of potential candidates

perfect retrieval distribution

$$\boxed{S} = \left\{ \boxed{b_i} : \max_{\forall \boxed{b_i} \in \boxed{B}} \boxed{P\left(\boxed{c_i}\boxed{= 1} \middle| \boxed{b_i}, \boxed{\theta}\right)} \right\}$$

candidate

true parameter

paragraphs (blobs)

candidate is data subject right

textual privacy policy

Sentence-based Information Retrieval

perfect retrieval distribution

$$S = \left\{ b_i : \max_{\forall b_i \in B} \boxed{P(c_i = 1 | b_i, \theta)} \right\}$$

$$\hat{f}(b_i, \hat{\theta})$$

extraction model

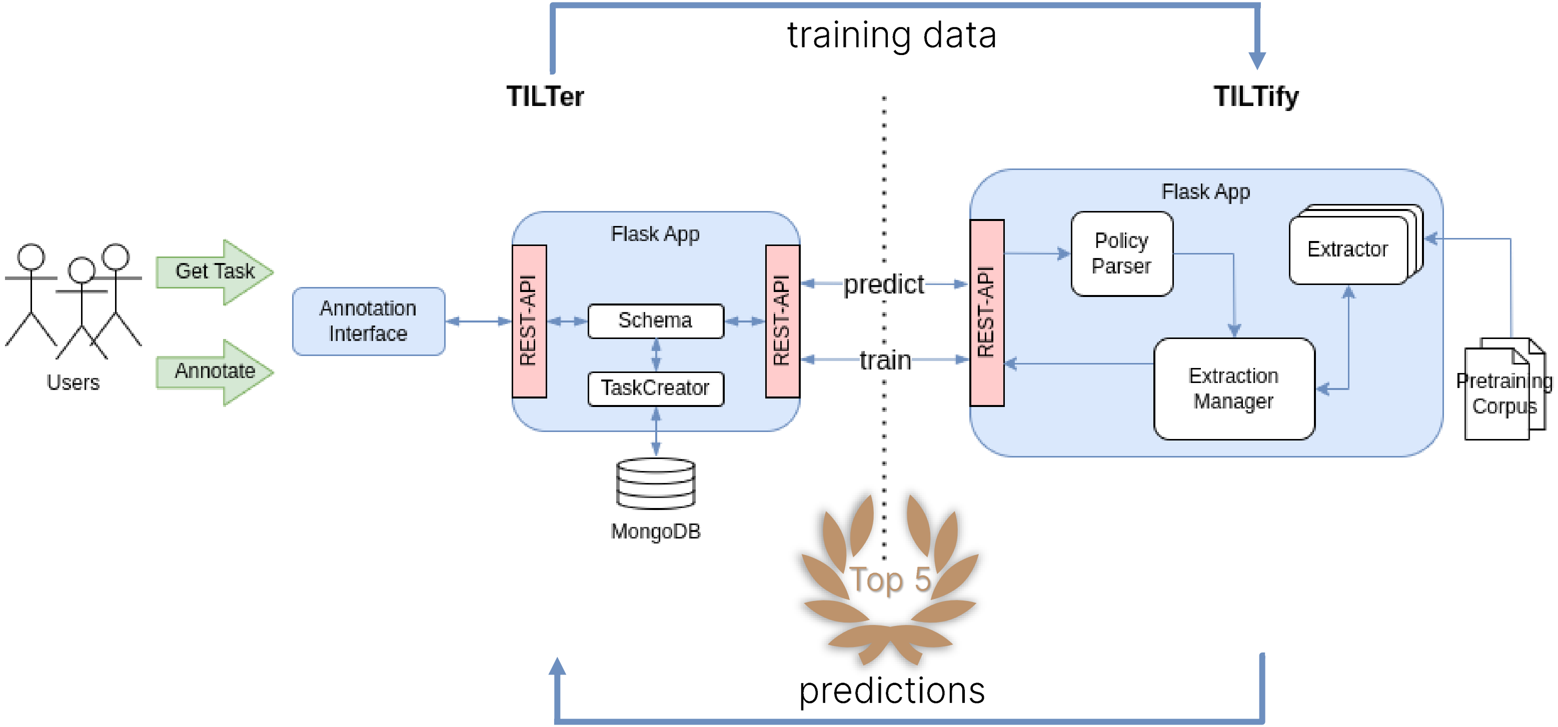
Extraction Model

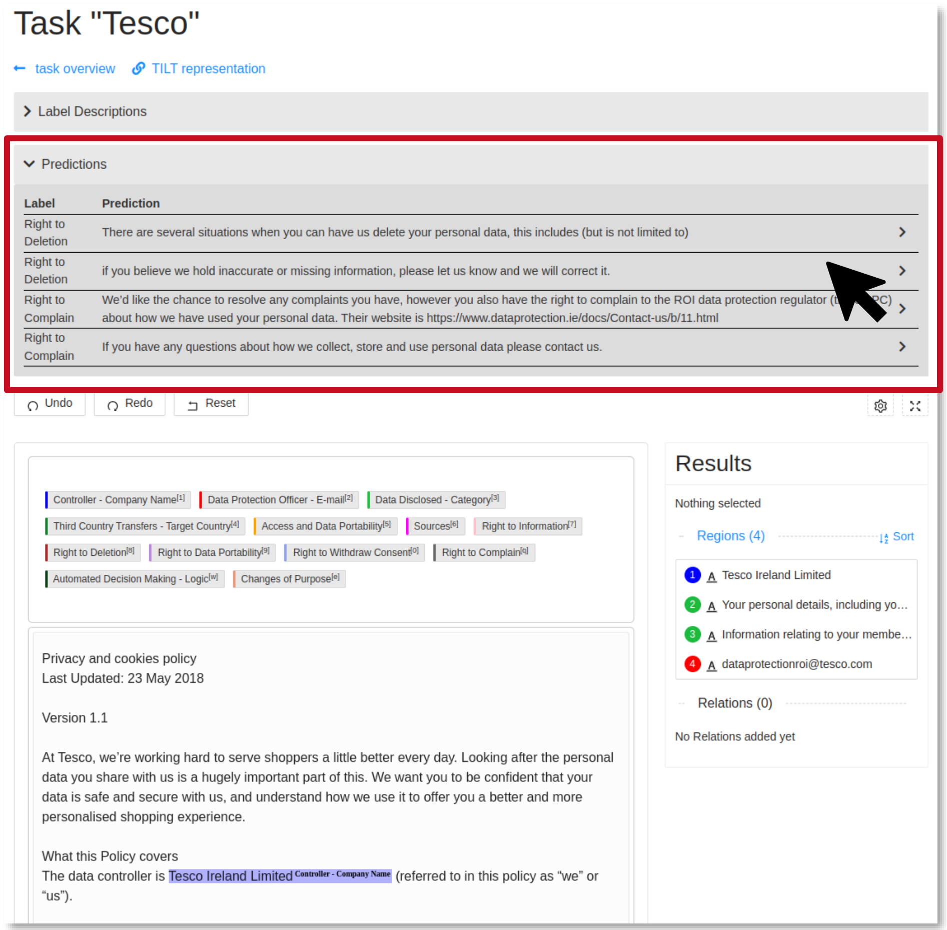
Static Word Embeddings & Naive Bayes

BinaryBERT

SentenceBERT

Human-in-the-Loop





Task "Tesco"

Label	Prediction	
Right to Deletion	There are several situations when you can have us delete your personal data, this includes (but is not limited to)	›
Right to Deletion	if you believe we hold inaccurate or missing information, please let us know and we will correct it.	›
Right to Complain	We'd like the chance to resolve any complaints you have, however you also have the right to complain to the ROI data protection regulator (t [illegible] PC) about how we have used your personal data. Their website is https://www.dataprotection.ie/docs/Contact-us/b/11.html	›
Right to Complain	If you have any questions about how we collect, store and use personal data please contact us.	›

Undo
Redo
Reset
Controller - Company Name[1]
Data Protection Officer - E-mail[2]
Data Disclosed - Category[3]
Third Country Transfers - Target Country[4]
Access and Data Portability[5]
Sources[6]
Right to Information[7]
Right to Deletion[8]
Right to Data Portability[9]
Right to Withdraw Consent[0]
Right to Complain[q]
Automated Decision Making - Logic[w]
Changes of Purpose[e]

Privacy and cookies policy
Last Updated: 23 May 2018

Version 1.1

At Tesco, we're working hard to serve shoppers a little better every day. Looking after the personal data you share with us is a hugely important part of this. We want you to be confident that your data is safe and secure with us, and understand how we use it to offer you a better and more personalised shopping experience.

What this Policy covers
The data controller is Tesco Ireland Limited Controller - Company Name (referred to in this policy as "we" or "us").

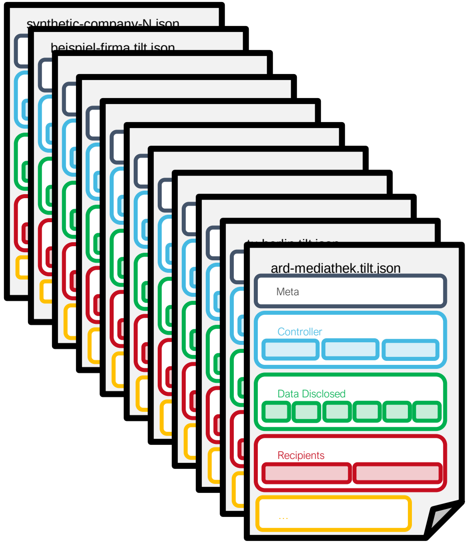
Results

Nothing selected

1 A Tesco Ireland Limited
2 A Your personal details, including yo…
3 A Information relating to your membe…
4 A dataprotectionroi@tesco.com

No Relations added yet

Evaluation
Dataset



60 annotated policies

- German language
- data scarcity
- differing amounts of data subject rights
- data imbalance:
  >16600 paragraphs
  <450 containing data subject rights

https://github.com/DaSKITA/tiltify/tree/main/data/annotated_policies

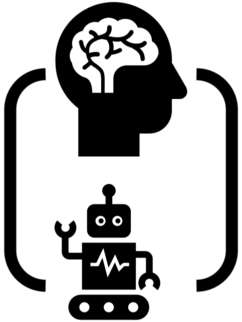
Evaluation
Models

	Static Word Embeddings & Naive Bayes	BinaryBERT	SentenceBERT
Right to Information	0.18	0.0	0.93
Right to Deletion	0.0	0.0	0.86
Right to Data Portability	0.0	0.0	0.86
Right to Complain	0.29	0.15	0.93
Right to Withdraw Consent	0.0	0.08	0.90

F1 Score (for 5-rank, if not stated otherwise)

Take-aways

We tackled problem of retrieving machine-readable transparency information from privacy policies under the constraint of data scarcity.



Adopted from https://datascientest.com/de/bert and https://tinyurl.com/2bhe3brk

S-BERT - in conjunction with k-rank approaches - is extremely promising to simplify the annotation of privacy policies, even with few training data.

A Human-in-the-Loop Approach for Information Extraction from Privacy Policies under Data Scarcity

Nicola Leschke

nl@ise.tu-berlin.de

tu.berlin/ise/nl

Nicola Leschke

DaSKITA

tu.berlin/ise/daskita

Gefördert durch:

Bundesministerium
für Umwelt, Naturschutz, nukleare Sicherheit
und Verbraucherschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

ptble
Projektträger Bundesanstalt
für Landwirtschaft und Ernährung