Engaging Content
Engaging People

Trinity College Dublin
The University of Dublin

Extracting Provenance Metadata from Privacy Policies

Harshvardhan J. Pandit, Declan O'Sullivan, Dave Lewis

ADAPT Centre, Trinity College Dublin, Dublin, Ireland
{ harshvardhan.pandit | declan.osullivan | dave.lewis **}** @ adaptcentre.ie

Identification

- Location in privacy policy
  - what sections does it occur under?
  - what is the context presented by the section?
- *Section 1* - data collection
- *Section 1.1* - data provided by user
- *Section 1.1.1* - data required for legitimate purposes
- "*Account Information*" - category of data
- "*sign-up"* data activity
- "*first name, last name*..." - types of data

Provenance Metadata Inference:
sign-up is a data collection activity that collects data of account information category directly from the user for legitimate purposes

Example Use-case: Airbnb Ireland
(no affiliation)

1. INFORMATION WE COLLECT
1.1 Information You Give to Us.
1.1.1 Information that is necessary for the use of the Airbnb Platform.
Account Information
When you sign up for an Airbnb Account, we require certain information such as your **first name**, **last name**, **email address**, and **date of birth**.

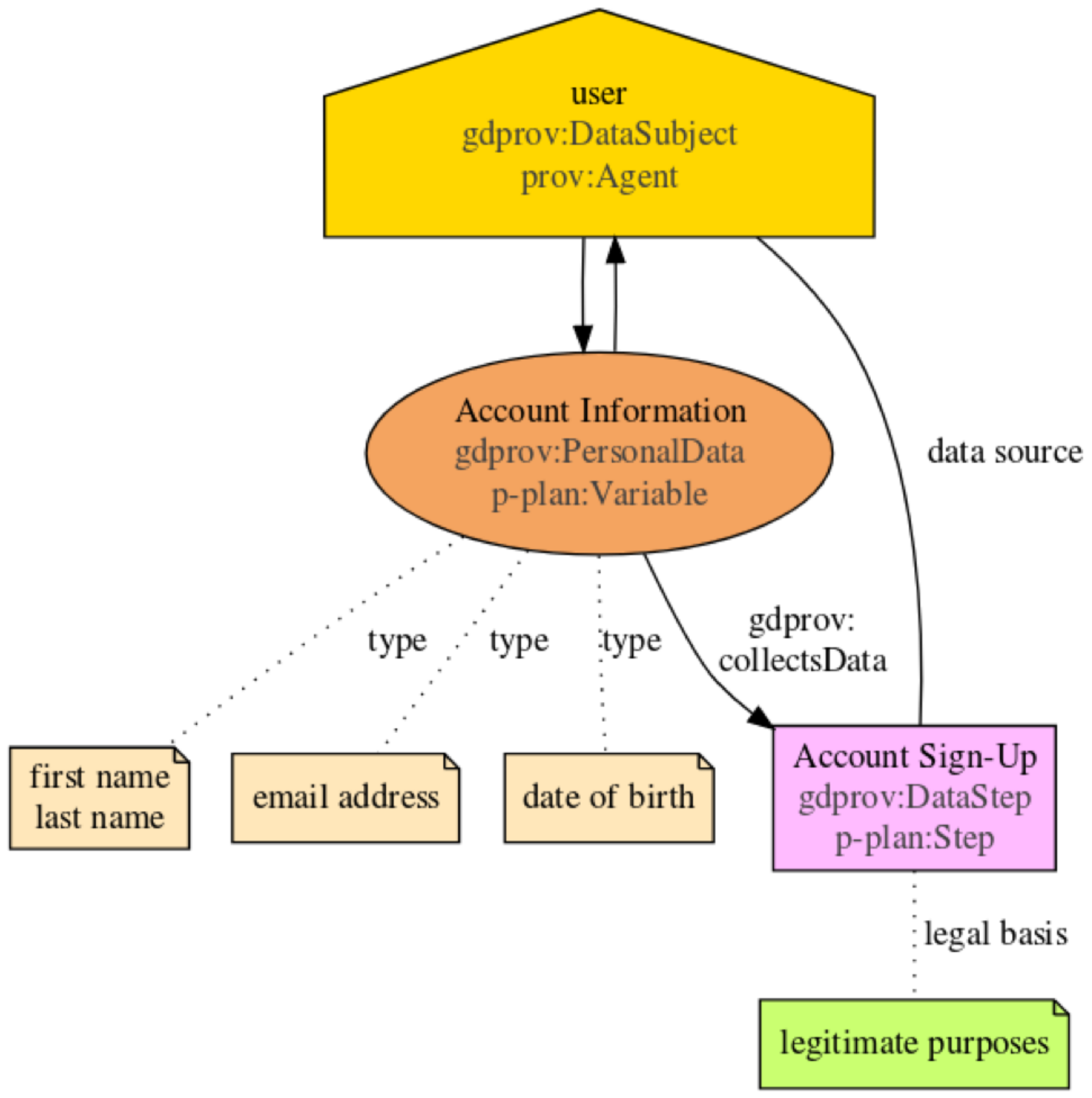
https://www.airbnb.ie/terms/privacy_policy
(accessed 16-APR-2018)

Extraction – I

- Use keyword-based Entity Recognition
- Search for keywords like *collect, share, store*
- Look for position in document to get context
- Extract metadata from surrounding sentences
- Terms and concepts from GDPRtEXT resource for GDPR-relevant keywords like *portability, breach*

Extraction – II

- Use Machine Learning to train algorithms
- Based on approach taken by UsablePrivacy Project to categorise statements in privacy policies based on expert annotations
- Sentence category provides information context
- Additional annotated corpus for training model



Representation

GDPRov ontology
- extends PROV-O and P-Plan
- defines terms using GDPR concepts
- 'model' an abstract representation or plan

Potential Applications

Visualise Privacy Policy
- represent data and actions over them
- easier to understand and comprehend
- accompany text to provide visual cues

Interpret Privacy Preferences
- Use provenance metadata to understand data actions
- Match with user's preferences (e.g. using ODRL)

Fondúireacht Eolaíochta Éireann
Science Foundation Ireland
Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.