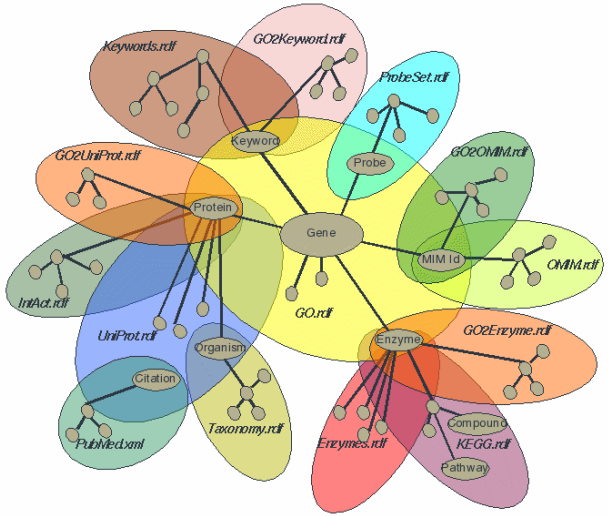
Extending the Environment Ontology with Text-mined Habitat Mentions

Riza Batista-Navarro (University of Manchester, UK)
Marie Angelique LaPorte (Biodiversity International, France)
Michael Regan (University of Colorado Boulder, USA)
William Ulate (Missouri Botanical Garden, USA)
Claus Weiland (BiK-F, Germany)

Background

- Ontologies
  - formal specification of domain-specific concepts and relations
  - crucial in knowledge representation, management and discovery
  - update and curation require human effort and time



Source: https://www.w3.org/2008/Talks/1009-bratt-W3C-SemTech

Environment Ontology

- initially developed to support the annotation of metagenomic data
- realigned its goals in support of the SDG Agenda for 2030
- much broader in scope (biodiversity and ecology)
- dramatic increase in number of classes

update and curation could benefit from automated support

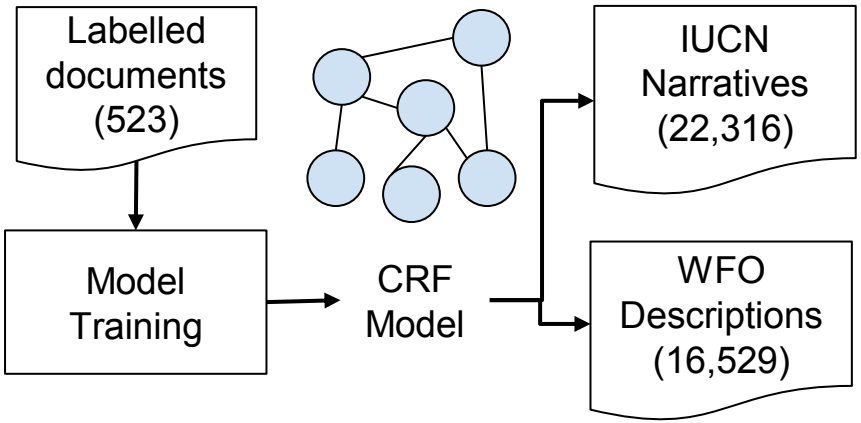
Overarching aim

- to help in expanding ENVO in a more efficient manner by automatically discovering new **habitat mentions**

Solution: Text mining

1. Extracting habitat mentions using NER
2. Text normalisation
3. Filtering
4. Clustering by semantic relatedness

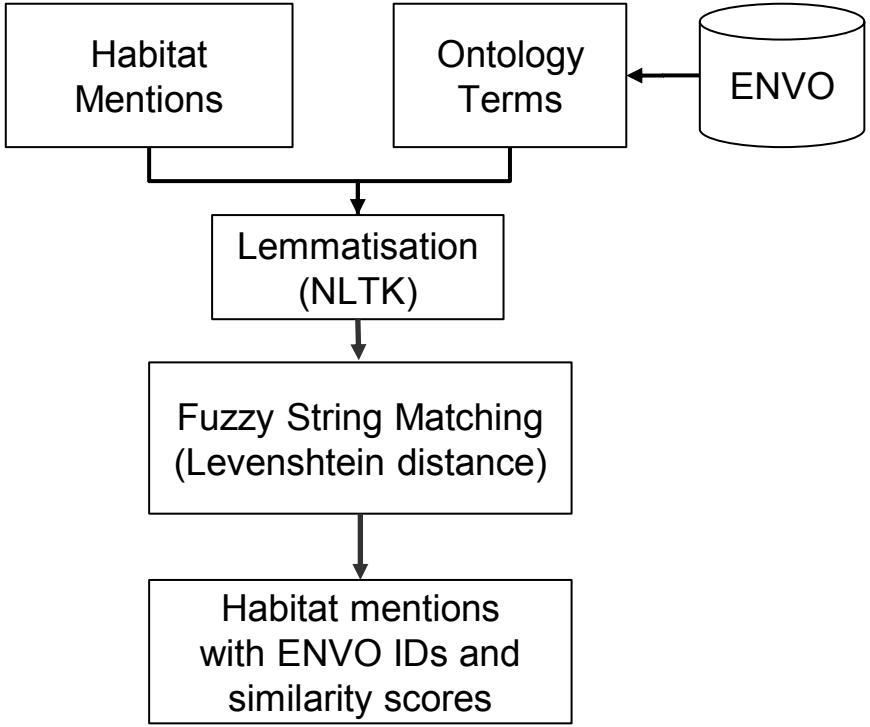
Extracting habitat mentions using NER



- Labelled documents (on forest trees): Biodiversity Heritage Library (BHL), grey literature, journals
- Model training
  - conditional random fields (CRFs)
  - features: character n-grams, word n-grams, morphology, dictionary matches (IUCN habitat terms)
- Result: 6,873 unique habitat mentions

IUCN = International Union for Conservation of Nature
WFO = World Flora Online

Text normalisation



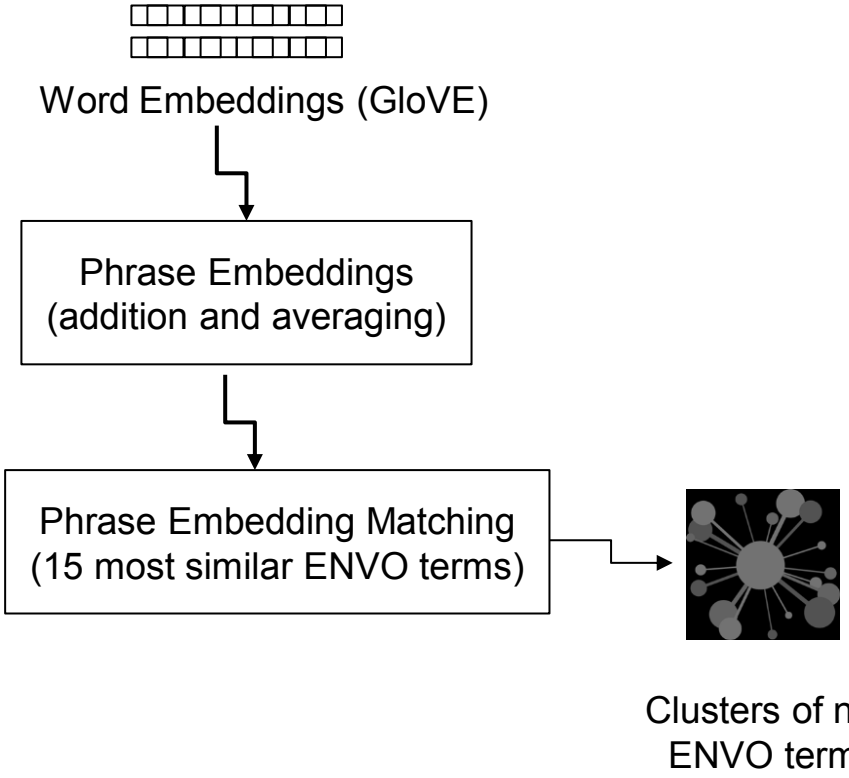
Filtering

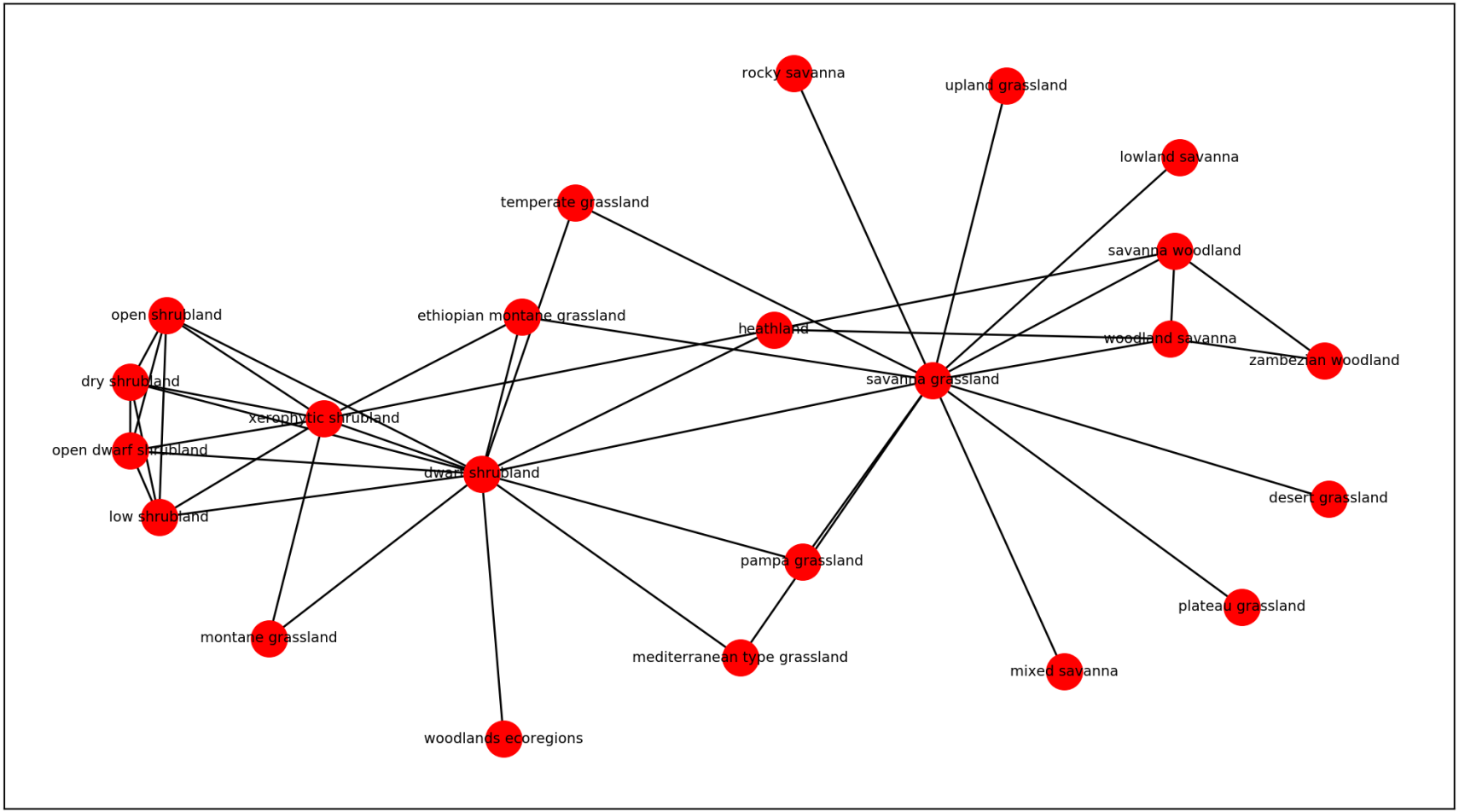
- thresholding on similarity score (score < 0.88)
- thresholding on frequency (frequency >=3)
- Result: 1,043 candidate new ENVO terms

Filtering (sample results)

Text-mined term	Most similar ENVO term	ENVO ID	Score	Freq
submontane forest	montane forest	ENVO_01000435	0.82	80
premontane forest	montane forest	ENVO_01000435	0.82	9
conifer forest	coniferous forest	ENVO_01000399	0.82	5
tropical moist forest	tropical rain forest	ENVO_00000109	0.81	15
tropical humid forest	tropical rain forest	ENVO_00000109	0.81	4
low deciduous forest	deciduous forest	ENVO_00000109	0.80	8
semideciduous forest	deciduous forest	ENVO_00000109	0.80	7

Clustering by semantic relatedness




rocky savanna
upland grassland
lowland savanna
temperate grassland
savanna woodland
open shrubland
ethiopian montane grassland
heathland
woodland savanna
zambezian woodland
dry shrubland
savanna grassland
xeric shrubland
open dwarf shrubland
dwarf shrubland
desert grassland
low shrubland
pampa grassland
plateau grassland
montane grassland
mediterranean type grassland
mixed savanna
woodlands ecoregions

Next steps

- Automatically submit new ENVO terms (GitHub)
- Validation by ENVO curators
- Expand scope (e.g., marine habitats)
- Link habitat mentions with geographic locations

Acknowledgement

- National Centre for Text Mining
- British Council (Newton Fund Institutional Links)
- National Science Foundation

Thank you!

Any questions?

Please contact:

Riza Batista-Navarro
School of Computer Science, University of Manchester
riza.batista@manchester.ac.uk