

eSTÓR and More:

Developing Irish language datasets
to combat language inequality

Dr Abigail Walsh, ADAPT Centre, Dublin City University

DCU

Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University



Engaging Content
Engaging People



An Roinn Turasóireachta, Cultúir,
Ealaíon, Gaeltachta, Spóirt agus Meán
Department of Tourism, Culture,
Arts, Gaelacht, Sport and Media



S20

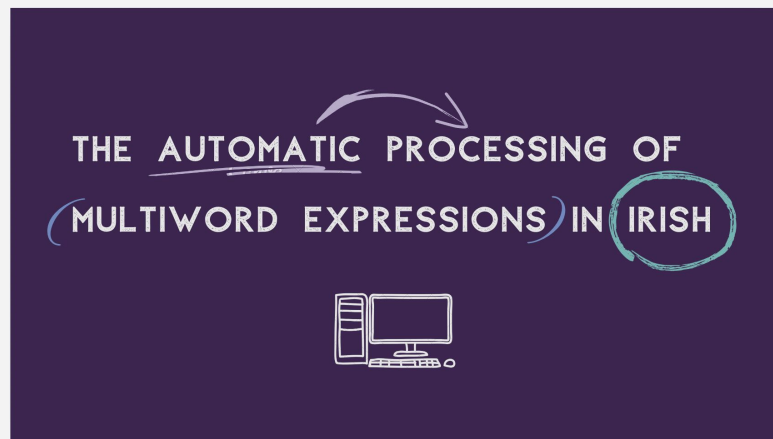
STRAITÉIS 20 BLIAIN DON
GHAELIGE 2010-2030
Ár dTeanga, Ár bPobal

About Me

Postdoctoral researcher at the
ADAPT Centre in Dublin City
University

Working with **Irish language data**
to improve **language technology**

Member of the eSTÓR team

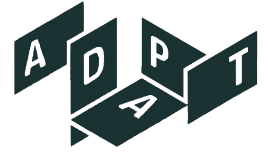




European Language Equality



European Language Equality



“Language Equality” EP Resolution (2018)

CELEX number: 52018IP0332

23.12.2019

EN

Official Journal of the European Union

C 433/42

P8_TA(2018)0332

Language equality in the digital age

European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI))

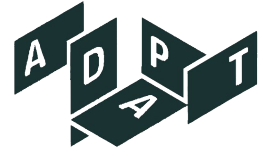
(2019/C 433/08)

The European Parliament,

- having regard to Articles 2 and 3(3) of the Treaty on the Functioning of the European Union (TFEU),
- having regard to Articles 21(1) and 22 of the Charter of Fundamental Rights of the European Union,

European Language Equality

ELE 1



<http://www.european-language-equality.eu>

Coordinator: DCU (ADAPT Centre)

Co-Coordinator: DFKI

Consortium: 52 partners (core partners: DCU – ADAPT Centre, DFKI, Charles University, ILSP, University of the Basque Country)

Objective: *Development of a strategic research, innovation and implementation agenda to achieve digital language equality in Europe by 2030*

Runtime: 18 months • January 2021 – June 2022



European Language Equality

ELE 1 + 2



- Main result: **Strategic Research and Innovation Agenda to achieve digital language equality in Europe by 2030**
 - Reports consolidating research from partners, SMEs and research networks

Follow-up Project ELE 2 ran for 12 months

- Additional feedback loop to **revise the SRIA**
- ELE 2 finished with the publishing of the ELE book
 - 33 language reports



Data Collection for European Languages

1. **Evidence-based investigation** of the level of technology support per language
2. **Aggregating** (meta)data through strategies including **manual input**
3. **European Language Grid Catalogue** served as **database** for further **DLE computations** (<https://live.european-language-grid.eu/catalogue/dashboard>)

Digital Language Equality (DLE)



Digital Language Equality (DLE) is the state of affairs in which **all languages** have the **technological support** and **situational context** necessary for them to continue to exist and to prosper as living languages in the digital age.

DLE Metric



The **DLE metric** is a measure that reflects the **digital readiness of a language** and its contribution to the state of **technology-enabled multilingualism**, tracking its progress towards the **goal of DLE**.

Technological Factors



Example: Corpora/Datasets

- Language(s)
- Domain(s)
- License
- Type of access
- **Corpus size**
- Etc.

<input type="checkbox"/>	_____
<input type="checkbox"/>	_____
<input type="checkbox"/>	_____
<input type="checkbox"/>	_____
<input type="checkbox"/>	_____

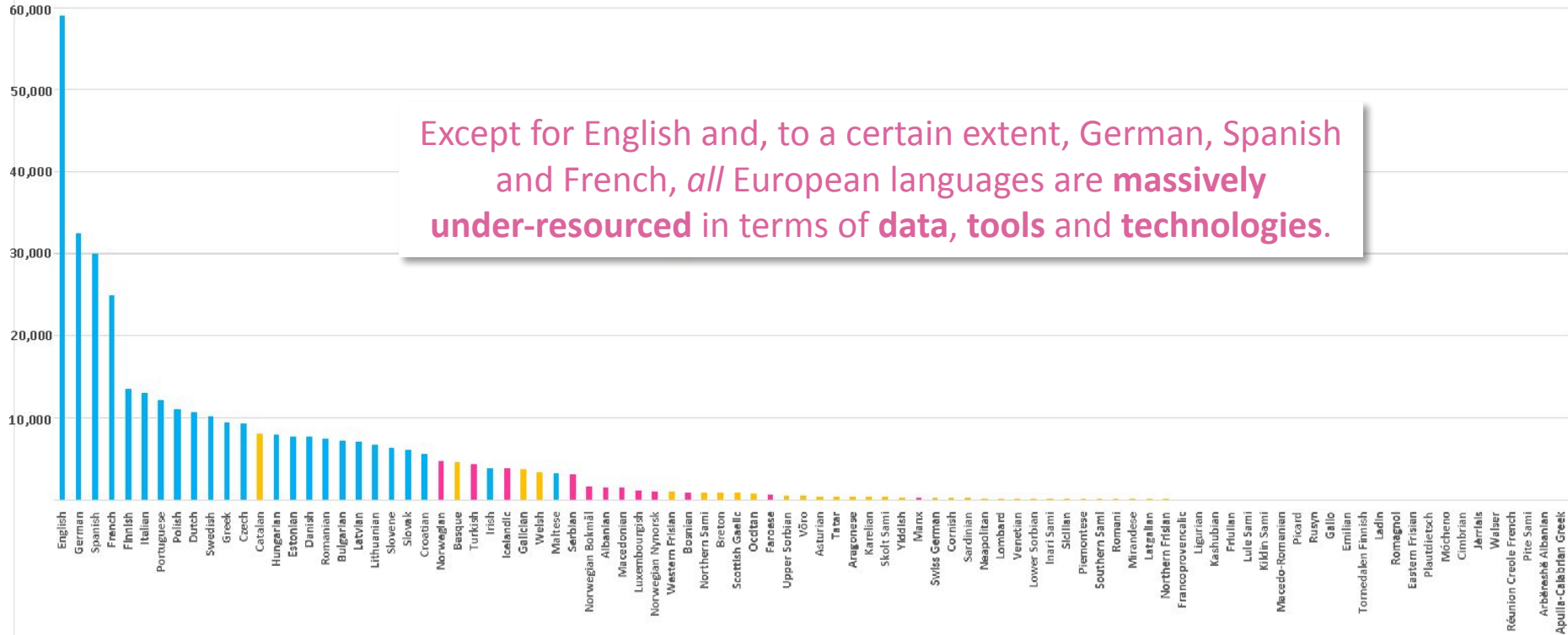
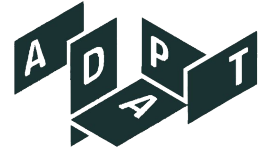
Contextual Factors



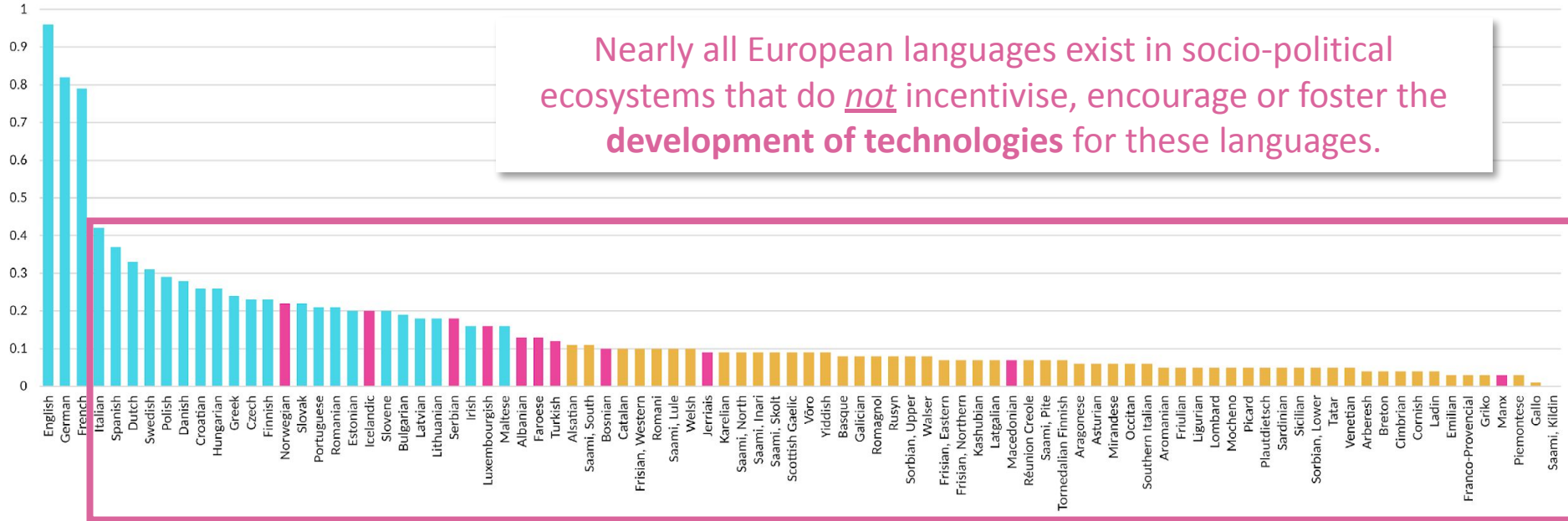
Example: Economy

- **Size** of economy of respective country, countries, region
- Size of **LT/NLP market** in country, countries, region
- Percentage of the **IT/ICT sector** relative to the economy
- **Funding programmes** targeting AI/LT/NLP start-ups
- Etc.

DLE Metric: Technological Scores



DLE Metric: Contextual Scores



Takeaway Message



- **AI Hype is real!**
 - Traction for Language Technology and NLP research
- All countries, sectors and languages interested in LLMs
- Coordination should happen on the **European level**
 - Solutions needed for the crucial challenges: e.g. data availability
 - Scientists are incentivised only to work on English—this needs to change!
 - National funding programmes needed to support technologies (LLMs) for their own languages

Takeaway Message




- Opportunities emerging
 - Horizon Europe (research) and Digital Europe (deployment)
- **Collaboration** and **coordination** between EU/EC and the different countries/regions
- **Language Data Spaces** (LDS) and the **Alliance for Language Technologies EDIC** (ALT-EDIC) will help bridge the **strategic plan for the EU** and the **participating countries**

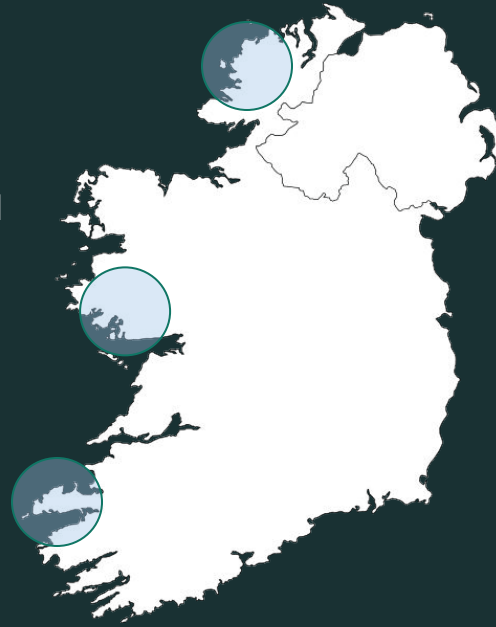


Irish Language: A Low-Resource Language

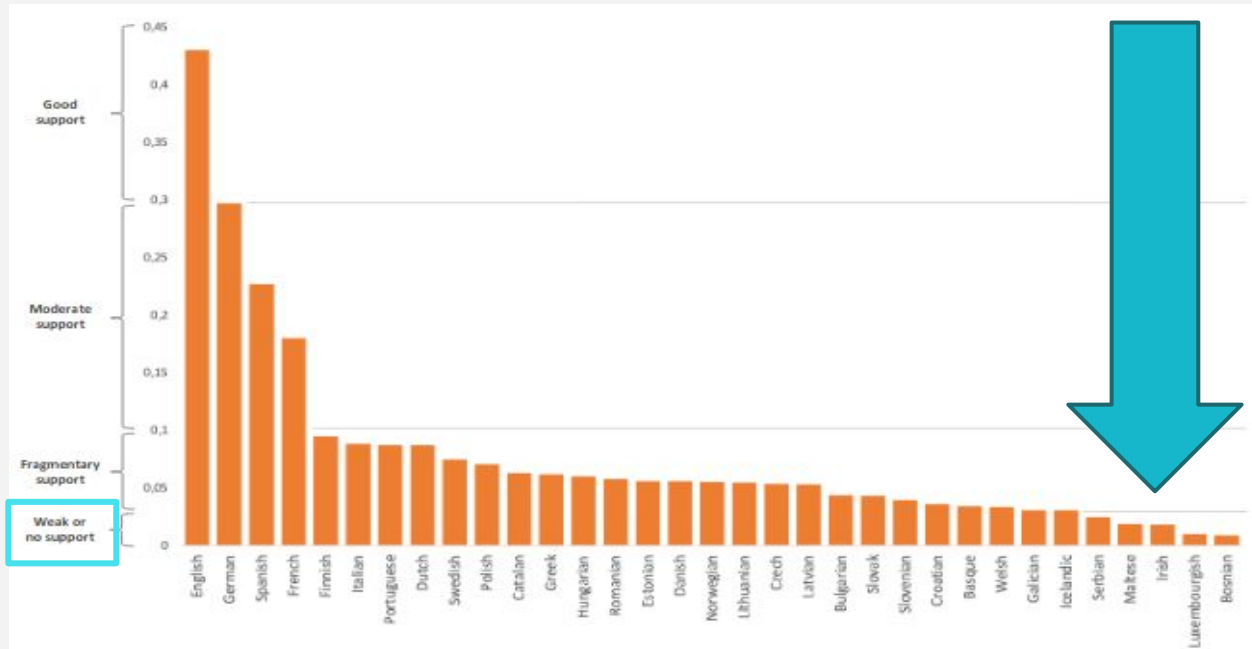
Irish Language Status



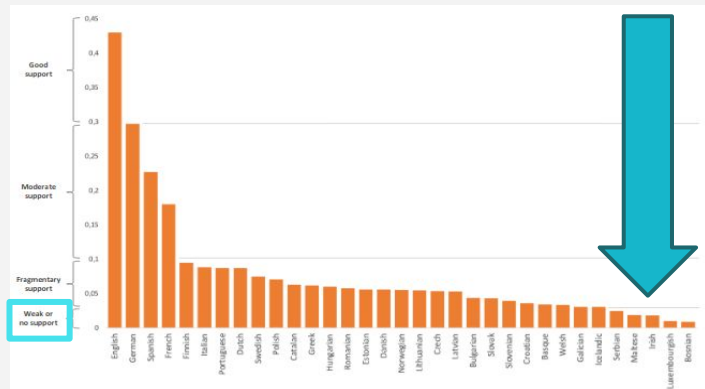
- Celtic language family
- First Official language of Ireland
 - National Language
 - Official Minority Language – Northern Ireland
 - Official EU Language
- Census (2022) : pop. 4,975,713
 - Ability to speak: 1,873,997  6%
 - Daily Usage: 71,968  1,835



Irish Language Status

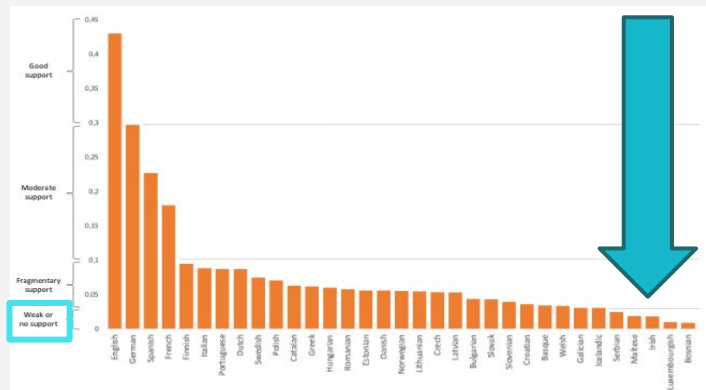


Irish Language Status



- **Digital Plan for Irish**
- Ireland's AI Strategy focuses on English LT
- Skill gap:
 - 1 undergraduate 3rd level course in computing and linguistics (with few Irish speaking students)
- Minimal support or investment from **industry**

Irish Language Status



- Still no
 - Question Answering Systems
 - Information Retrieval
 - Virtual Agents/chatbots
 - Adaptive Learning

Recommendations: ELE Report on Irish



- **Change of focus**
 - Away from current focus on dictionary development and training translators
- **Untapped potential for machine learning/ data-driven technologies**
 - Increase awareness: **value** of language data
 - Improve: language data **management practices**
 - Increase awareness: importance of **open** data (sharing)
- Same for other European languages (e.g. Norwegian)





Introducing

eSTÓR

European Language Resource Infrastructure (ELRI) project



- Language Resources shared with the **European Commission**
 - Improve **translation services** in Europe
 - Co-funded by the European Union
- **National Relay Stations (NRS)** set up in **Spain, France, Portugal, and Ireland**
 - Gather, process and share

PRINCIPLE Project



- Creation of bespoke MT engines for public administration
 - **Iceland, Norway, Croatia** and **Ireland**



An Roinn Dlí agus Cirt
agus Comhionannais
Department of Justice
and Equality

PRINCIPLE



Foras na Gaeilge

eSTÓR 2021-2023



Relaunching NRS as eSTÓR

- Completely rewritten site
- Modern language and framework
- New toolchain
- 3 year funding (2023 – 2025) from
Department of Tourism, Culture, Arts,
Gaeltacht, Sport and Media



A decorative graphic consisting of numerous thin, light blue lines that flow and wave across the dark teal background of the header section.

ESTÓR

Sonraí Teanga Óstáilte i gcomhair Ríomhphróiseála

283

Datasets

62

Publishers

WELCOME TO ESTÓR IN IRELAND!

This is the eSTÓR site in Ireland, where language resources can be collected, prepared and shared amongst public institutions and translation centres.

JOIN ESTÓR

Members of public organizations in Ireland can join to download shared resources, contribute their own data and benefit from translation memories or other language resources automatically prepared by the eSTÓR engines.

eSTÓR: **Sonraí Teanga Óstáilte** i gcomhair **Ríomhphróiseála**

- **Maintenance and improvements** to eSTÓR website
- **Outreach activities** to increase number of users and uploads of data to site
- **Digitisation** of Statutory Instruments and Acts in the Rannóg an Aistriúcháin
- **Development of tools** to extract Irish language data from PDF files



Ornait O'Connell
Language Officer



Jane Dunne
Project Manager



Teresa Clifford
PhD Student



Dr Brian Davis
PI



Mark Andrade
Research Assistant



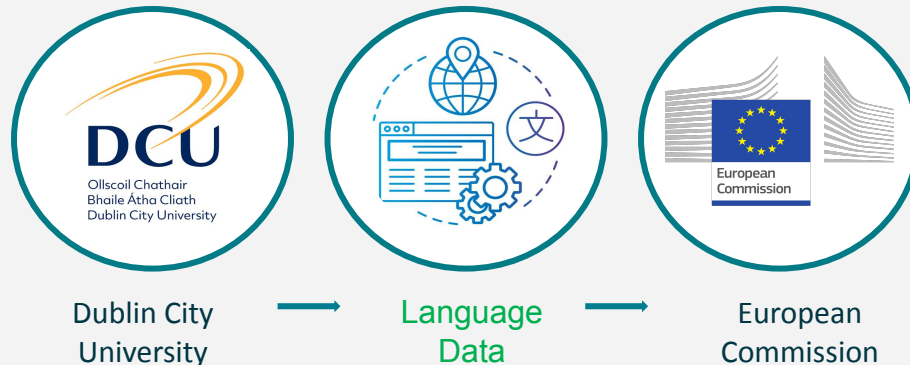
Dr Abigail Walsh
Postdoc Researcher



Órla Ní Loinsigh
Software Engineer

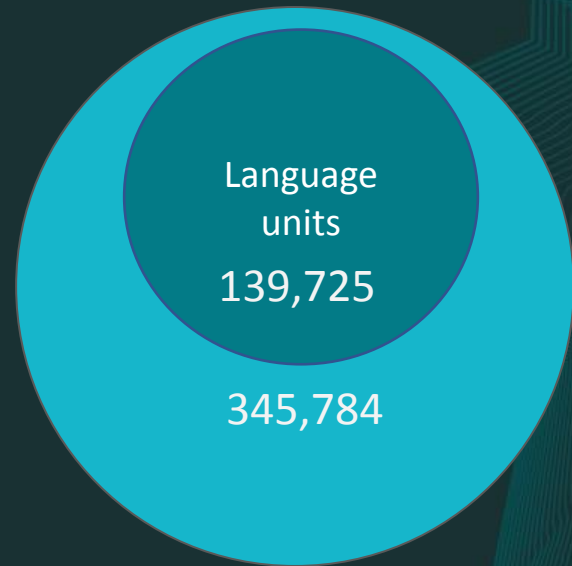
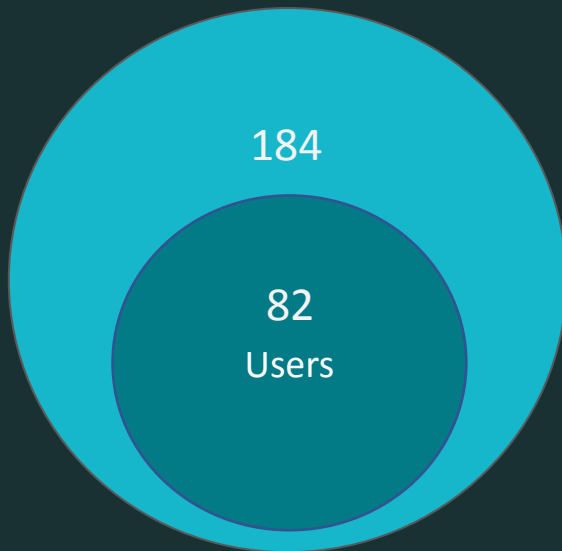
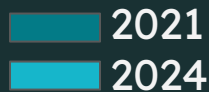
eSTÓR: Sonraí Teanga Óstáilte i gcomhair Ríomhphróiseála

- Both bilingual English-Irish and monolingual Irish text data
- Online platform allows for uploading data easily
 - Multiple formats accepted





eSTÓR Data Collection to Date



eSTÓR Processing



(.doc, .docx,
.odt, .rtf)



(.pdf)



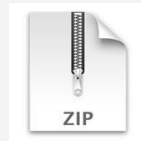
(.txt, .xml, .tbx)



(.xls, .xlsx)



(.tmx, .sdlTM)



(.zip)

Aims

- Creating clean, **aligned** data from raw input data
- Automating as far as possible

2 [239]

S.I. No. 239 of 2013

SUPREME COURT AND HIGH COURT (FEES) ORDER 2013

1, ALAN SHATTER, Minister for Justice and Equality, in exercise of the powers conferred on me by section 65 (as amended by section 66 of the Civil Law (Miscellaneous Provisions) Act 2011 (No. 23 of 2011)) of the Courts of Justice Act 1936 (No. 48 of 1936) (as adapted by the Justice and Law Reform (Alteration of Name of Department and Title of Minister) Order 2011 (S.I. No. 138 of 2011)), with the consent of the Minister for Public Expenditure and Reform, hereby order as follows:

1. (1) This Order may be cited as the Supreme Court and High Court (Fees) Order 2013.

(2) This Order comes into operation on 10 July, 2013.

2. In this Order—

“Act of 1988” means the Bankruptcy Act 1988 (No. 27 of 1988);

2 [239]

I.R. Uimh. 239 de 2013

AN tORDÚ UM AN gCÚIRT UACHTARACH AGUS AN ARD-CHÚIRT (TÁILLÍ), 2013

I bhfeidhmiú na gcumhachtaí a thugtar dom le halt 65 (arna leasú le halt 66 den Acht um an Dlí Sibhialta (Forálacha Ilghnéitheacha), 2011 (Uimh. 23 de 2011)) den Acht Cúirteanna Breithiúnais, 1936 (Uimh. 48 de 1936) (ama oiriúnú leis an Ordú Dlí agus Cirt agus Athchóirithe Dlí (Aim na Roinne agus Teideal an Aire a Athrú), 2011 (I.R. Uimh. 138 de 2011)), le toiliú an Aire Cateachais Phoiblí agus Athchóirithe, ordaímse, ALAN SHATTER, Aire Dlí agus Cirt agus Comhionannais, leis seo, mar seo a leanas:

1. (1) Féadfar an tOrdú um an gCúirt Uachtarach agus an Ard-Chúirt (Táillí), 2013 a ghairm den Ordú seo.

(2) Tíocfaidh an tOrdú seo i ngníomh an 10 Iúil 2013.

2. San Ordú seo—

ciallaíonn “Acht 1988” an tAcht Féimheachta, 1988 (Uimh. 27 de 1988);

```
<tu tuid="6">
  <tuv xml:lang="en-IE">
    <seg>I, ALAN SHATTER, Minister for Justice and Equalit
    seg>
  </tuv>
  <tuv xml:lang="ga-IE">
    <seg>I bhfeidhmiú na gcumhachtaí a thugtar dom le halt
    seo, mar seo a leanas:</seg>
  </tuv>
</tu>
<tu tuid="7">
  <tuv xml:lang="en-IE">
    <seg>1. (1) This Order may be cited as the Supreme Cou
    </tuv>
  <tuv xml:lang="ga-IE">
    <seg>1. (1) Féadfar an tOrdú um an gCúirt Uachtarach a
    </tuv>
</tu>
<tu tuid="8">
  <tuv xml:lang="en-IE">
    <seg>(2) This Order comes into operation on 10 July, 2
    </tuv>
  <tuv xml:lang="ga-IE">
    <seg>(2) Tíocfaidh an tOrdú seo i ngníomh an 10 Iúil 2
    </tuv>
</tu>
<tu tuid="9">
  <tuv xml:lang="en-IE">
    <seg>2. In this Order</seg>
  </tuv>
  <tuv xml:lang="ga-IE">
    <seg>2. San Ordú seo</seg>
  </tuv>
</tu>
<tu tuid="10">
  <tuv xml:lang="en-IE">
    <seg>"Act of 1988" means the Bankruptcy Act 1988 (No.
    </tuv>
  <tuv xml:lang="ga-IE">
```


Challenges

- Legacy Data
 - OCR and Irish (including English that features Irish terms/proper nouns)
 - Repeated file conversion and corruption
- PDFs: Sentence boundary information and table data
- Boilerplate/headers/footers
- Invisible Errors
- Verification



Image from Glen Noble from Unsplash

eSTÓR Conclusions




- Government support is *crucial*
 - Data sharing policies, funding, digital strategy etc.
- EU Support is *crucial*
- Education and broad outreach is *crucial*
- **More data leads to further development**
 - Positive feedback loop

Go Raibh Maith Agaibh!
Thank you!
Tusen Takk!

<https://estor.ie/> (Only Accessible in Ireland)

abigail.walsh@adaptcentre.ie 

@estór_eireann 

eSTÓR na hÉireann 

European Language Grid Dashboard

