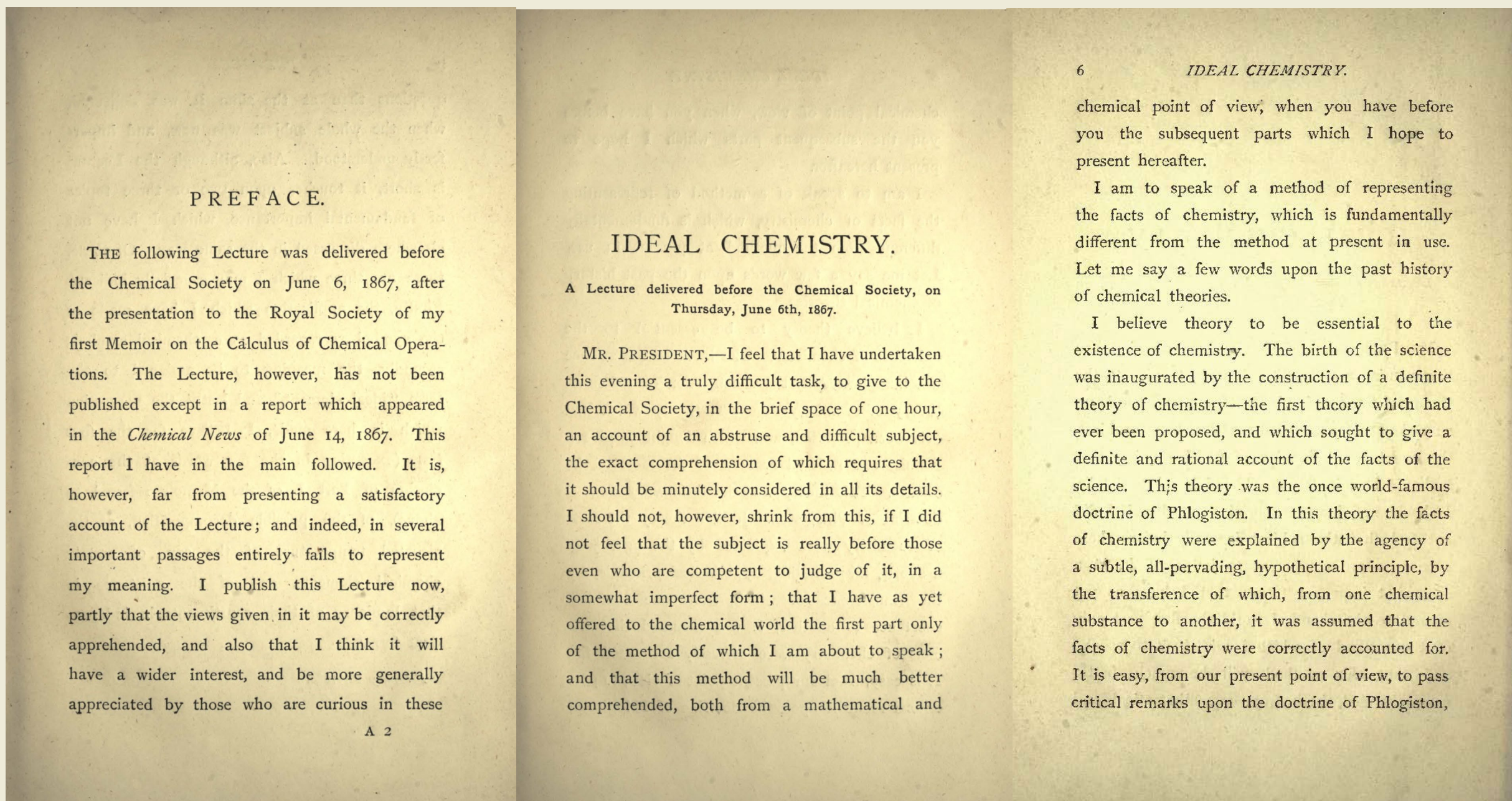







Do texto histórico ao computador: metodoloxía e criterios de recompilación do Coruña Corpus









CARACTERÍSTICAS E FINALIDADE

-  Corpus a gran escala
-  Diacrónico
-  Especializado.
-  Mostras de inglés científico (1700-1900).
-  Finalidade: investigar a evolución do inglés científico no período do inglés moderno tardío (1700-1900)

CRITERIOS DE RECOPIACIÓN DAS MOSTRAS





Representatividade

-  Só textos en prosa editados e impresos
-  Preferencia por primeiras edicións ou outras publicadas dentro dos 30 anos seguintes (Kytö, Rudanko & Smitterberg, 2000: 92)
-  Control de variación
-  Para evitar idiosincrasias de autor:
→ 1 mostra por autor en todo o corpus
-  Para evitar interferencias lingüísticas:
→ Só autores educados en inglés
→  Sen traducións (nin sequera realizadas polos propios autores)

Equilibrio

- ✖ Sen prólogos nin dedicatorias
- 📄 Extracción de fragmentos de diferentes partes dos textos
- 📏 Para textos de menos de 10.000 palabras:
 - inclusión completa (*in toto*)
 - 📏 Como consecuencia, algunhas décadas conteñen 3 mostras en lugar de 2

- ## Parámetros extralingüísticos para garantizar equilibrio

-  Sexo
-  Lugar de formación
-  Xénero textual
-  Contexto da disciplina no período histórico



DESEÑO E ESTRUTURA

Modular: subcorpus adicados a disciplinas científicas individuais (astronomía, filosofía, historia, física, etc),

punto de partida

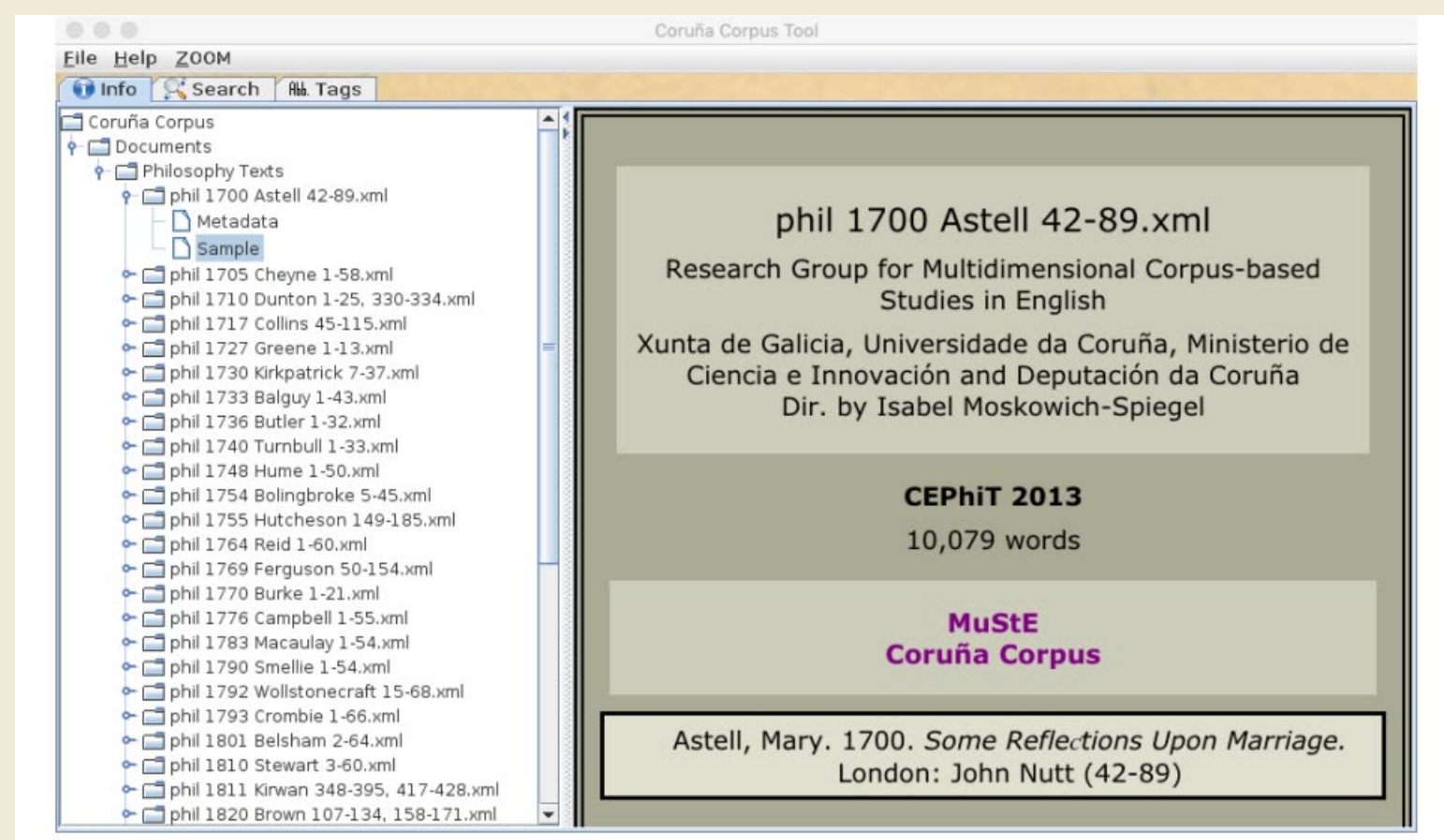


a clasificación UNESCO (1978) da ciencia
(adaptada ás epistemoloxías históricas)

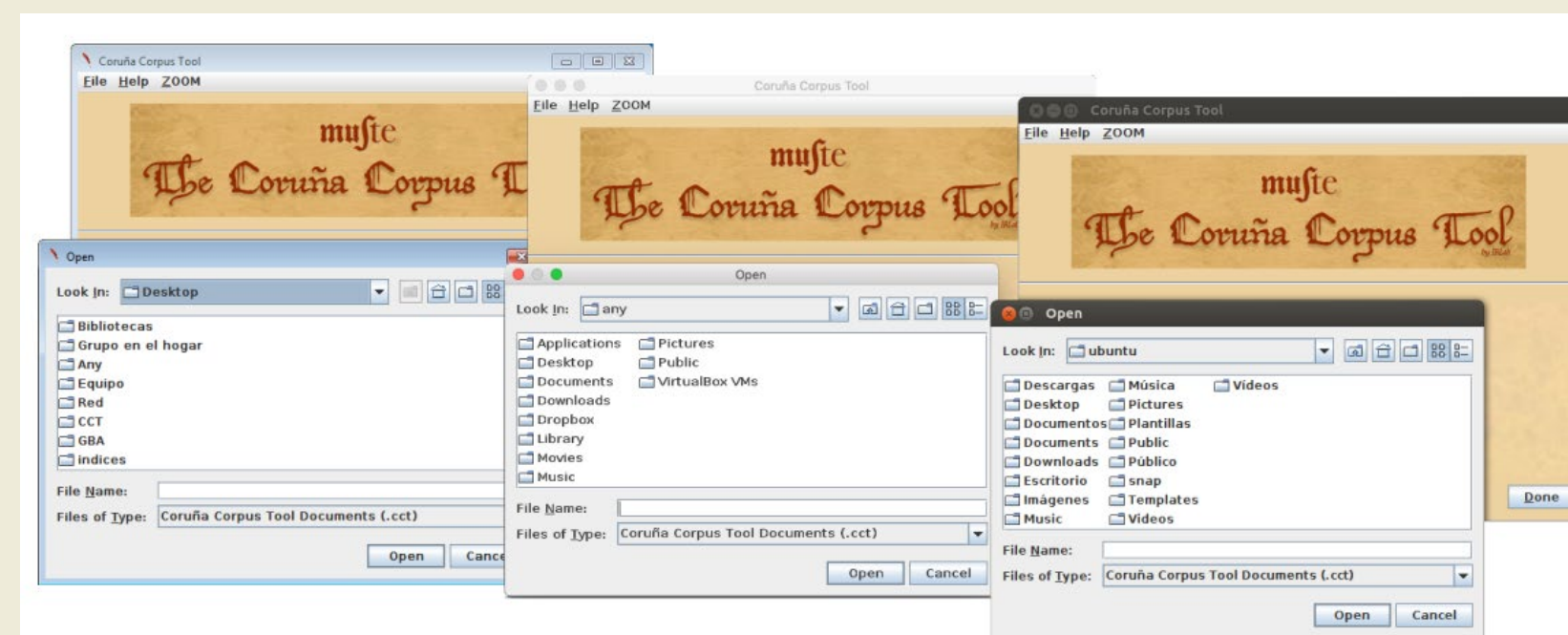
Tamaño do corpus

- Aproximadamente 10.000 palavras por mostra
- 2 mostras por década e disciplina
- Aproximadamente 200.000 palavras por século e disciplina
- Aproximadamente 400.000 palavras por subcorpus

TRATAMIENTO E CODIFICACIÓN DE DATOS



FERRAMENTA Coruña Corpus Tool (CCT)



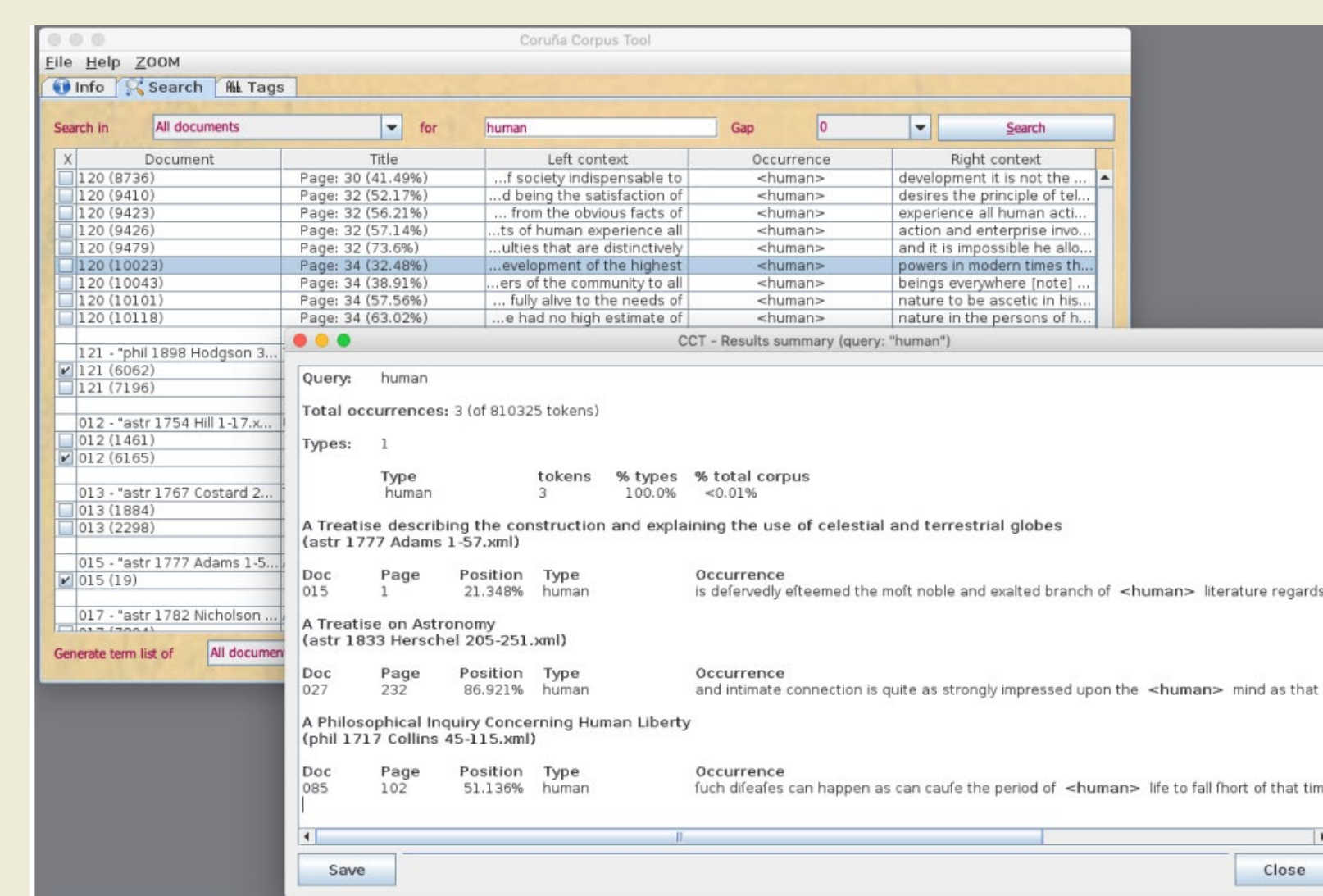
- Busca avanzada
- Concordancias (KWIC)
- Análisis cuantitativa e cualitativa



Subcorpus	N. palabras
CETA	409909
CHET	404424
CEPhiT	401129
CElIST	400305
Total	1615767

- Codificación XML-TEI: Etiquetaxe XML consonte ás directrices da TEI (Text Encoding Initiative)

- Transcrición manual dos textos orixinais
- **3** revisións manuais para garantir unha representación fiel do texto fonte
- Exclusión de táboas, figuras, fórmulas, gráficos e citas
- Representación de graffías antigas e caracteres especiais con Unicode: (l), ligadura (ct), signos do zodíaco, etc.
- Inclusión dun ficheiro de metadatos en XML asociado a cada mostra textual



METADATOS



Texto

☐ Xénero textual

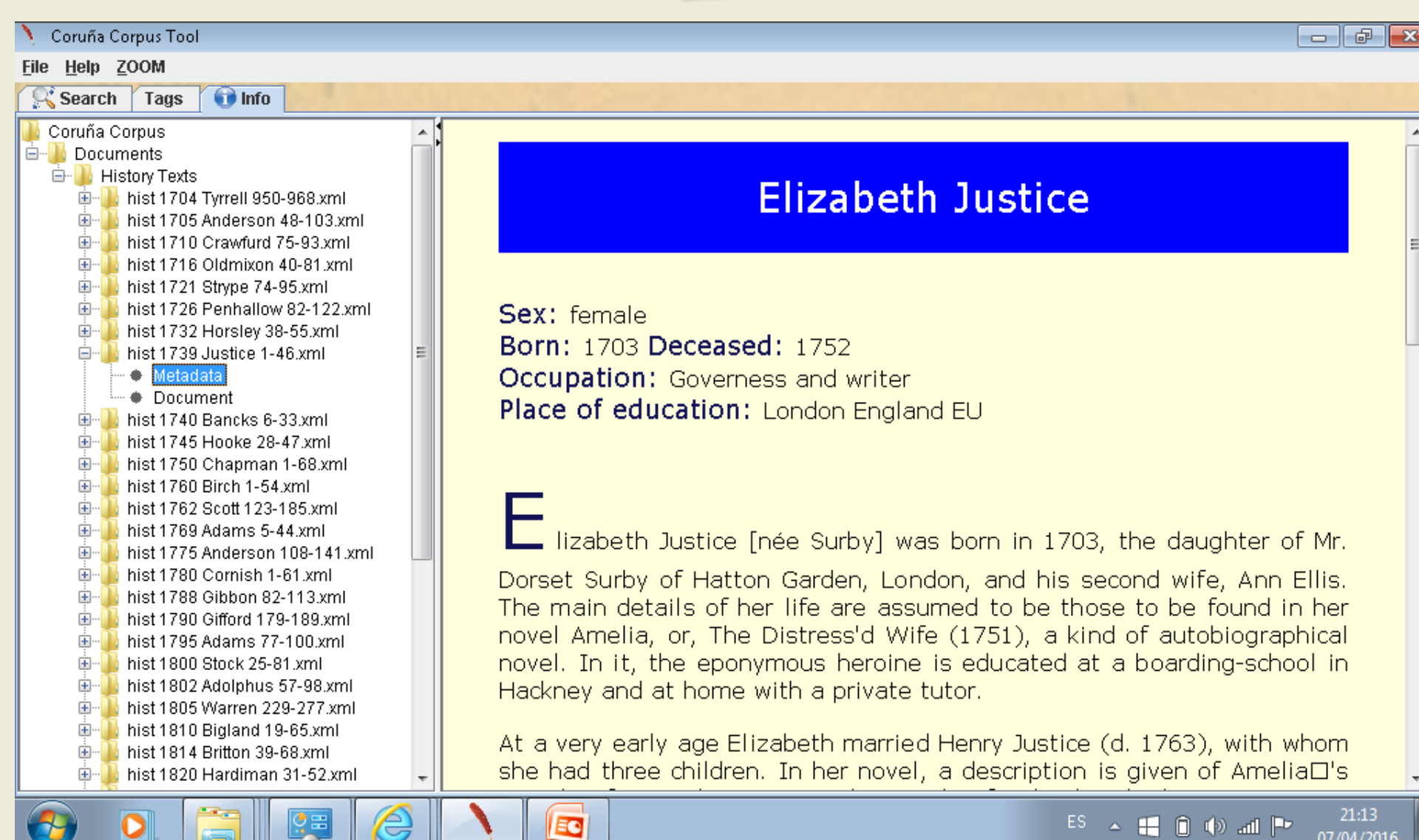
☐ Ano, lugar publicación

Author/a

  **Sexo**

☐  Procedencia xeográfica

 Idade



ALGUNHAS PUBLICAÇÕES

<https://www.udc.es/grupos/muste/index.html>

Barsaglini-Castro, Anabella and Valcarce, Daniel. 2020. The Coruña Corpus Tool: Ten Years On. *Revista de Procesamiento del Lenguaje Natural*, 64, 12-19.

Biber, D. 1993. Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257. <https://doi.org/10.1093/lc/8.4.243>

Crespo, Begoña & Isabel Moskowich. 2010. [“CETA in the Context of the Coruña Corpus”](#). *Literary and Linguistic Computing*, 25(2): 153–164.

Crespo, Begoña & Isabel Moskowich. 2020. [Astronomy, Philosophy, Life Sciences and History Texts: Setting the Scene for the Study of Modern Scientific Writing](#). *English Studies*.

al. (eds.) 'Of Varying Language and Opposing Creed': New Insights into Late Modern English. Bern: Peter Lang, 341–357.