Thoughts on data sustainability

John Coleman

Phonetics Laboratory,
University of Oxford

Outline

- Why speech corpora are important
- What we did in Digging into Data round 1
- What we didn’t do in Digging into Data round 1 that we sort of thought we might do (but were careful not to promise to do)
- What we’ve done since Digging into Data round 1
- A few thoughts about today’s projects
- Sustainability

UNIVERSITY OF
OXFORD

Why speech corpora are important

- Speech is (still) a fair proportion of the largest flow of data in the world (video & telephony)
- Languages are enormous, and everyone talks somewhat differently from anyone else
- Variation is everywhere, so very large amounts of data are needed in order to make statistically reasonable inferences (much linguistics research does not)
- Zipf's law[1] (lopsided sparsity) is a killer

[1] Technically, it's a Yule distribution

DiD 1 “Mining a Year of Speech”

- UK side: Coleman, Kochanski, Ravary and Burnard at University of Oxford Phonetics Laboratory
- Jonnie Robinson and colleagues at British Library
- US side: Mark Liberman, Jiahong Yuan and colleagues at UPenn Linguistic Data Consortium

UNIVERSITY OF OXFORD

What we did in DiD Round 1

- Drew together a number of large, transcribed, accessible corpora of UK and US English
- Force-aligned the transcriptions to the audio
- Openly published the UK audio and alignments at http://www.phon.ox.ac.uk/AudioBNC

What we did in DiD Round 1

- Used HTML5 syntax for audio fragments to allow users to access *any* part of the recordings, not just transcribed speech
- e.g. http://bnc.phon.ox.ac.uk/data/021A-C0897X0229XX-ABZZP0.wav?t=1560,1582

  (play audio)

UNIVERSITY OF OXFORD

What we didn’t do in DiD Round 1

- Make a user-interface for searches etc
- Devise/promote a new XML syntax for spoken audio in language transcriptions

What we’ve done since DiD Round 1

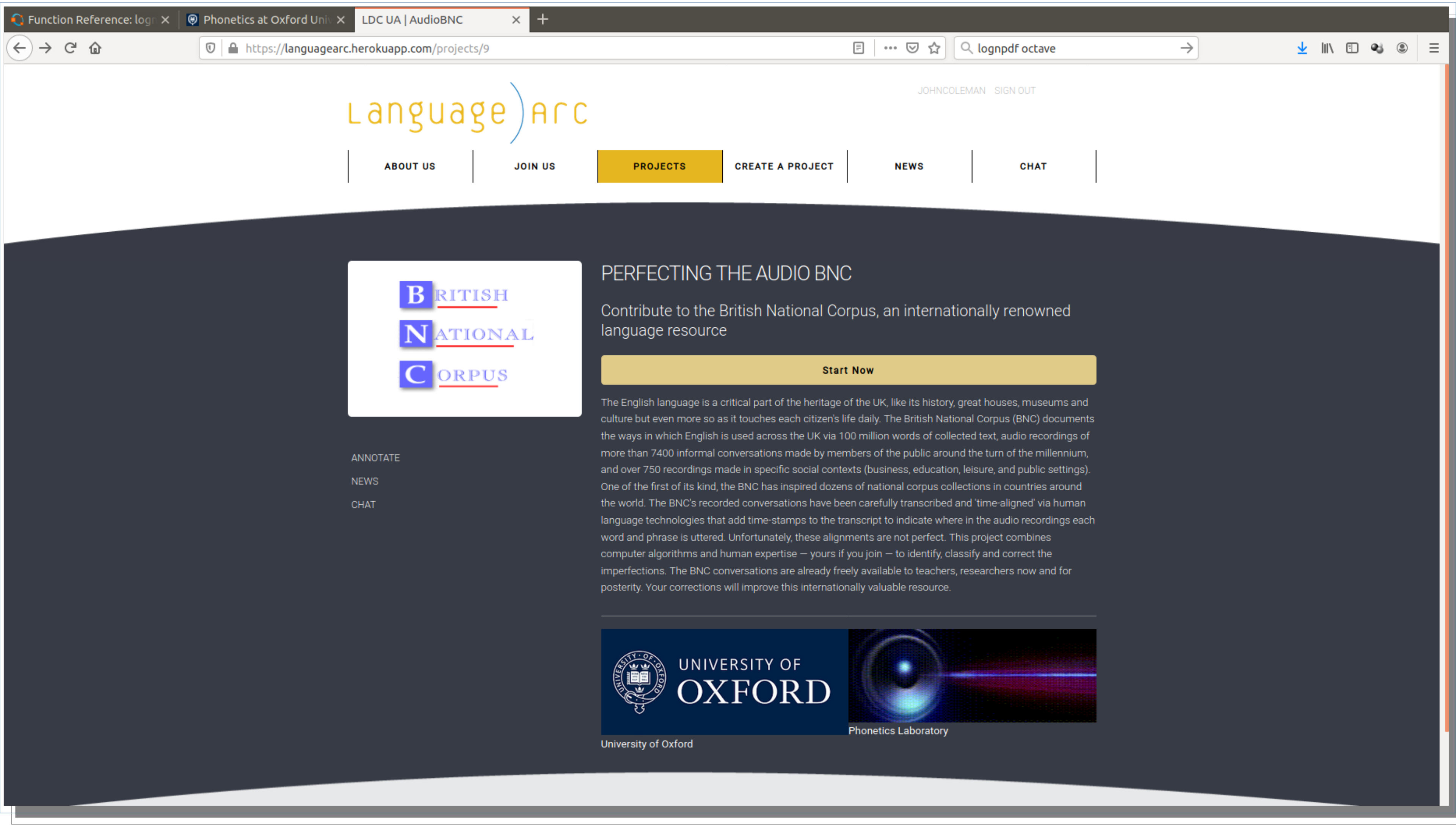
- Big project looking at phonetic details of word-joins in UK English
- Contributed to a huge study of the usage of “um” and “er” in English and other Germanic languages
- A lot of student projects
- Applied for and failed repeatedly to get funding to do more work on federating spoken English corpora (so hurrah for SPADE)

What we’ve done since DiD Round 1

- Helped other sites link to the audio. e.g.

  http://corpora.lancs.ac.uk/BNCweb/

  so we didn’t have to build a search tool, user interface etc
- New LDC project to crowd-source corrections to AudioBNC alignments: LanguageARC.org, launching in May



PERFECTING THE AUDIO BNC

Contribute to the British National Corpus, an internationally renowned language resource

The English language is a critical part of the heritage of the UK, like its history, great houses, museums and culture but even more so as it touches each citizen's life daily. The British National Corpus (BNC) documents the ways in which English is used across the UK via 100 million words of collected text, audio recordings of more than 7400 informal conversations made by members of the public around the turn of the millennium, and over 750 recordings made in specific social contexts (business, education, leisure, and public settings). One of the first of its kind, the BNC has inspired dozens of national corpus collections in countries around the world. The BNC's recorded conversations have been carefully transcribed and 'time-aligned' via human language technologies that add time-stamps to the transcript to indicate where in the audio recordings each word and phrase is uttered. Unfortunately, these alignments are not perfect. This project combines computer algorithms and human expertise — yours if you join — to identify, classify and correct the imperfections. The BNC conversations are already freely available to teachers, researchers now and for posterity. Your corrections will improve this internationally valuable resource.

University of Oxford

Phonetics Laboratory

UNIVERSITY OF OXFORD

A few thoughts about today’s projects

- *SPADE*
- Succeeds in drawing multiple speech corpora together, by aggregation not federation
- Facilitates some nice big investigations, e.g. *bead* vs *beat*: 1964 speakers, 30 dialects, ~230k examples

A few thoughts about today’s projects

- *SPADE*
- Aggregation: harder to build up in a community than federation
- Uses a permissions system to protect data and observe legal constraints

A few thoughts about today’s projects

- *ACLEW*
- Daylong audio recordings
- Time-aligned annotations, (mostly/entirely?) automatic
- “Sharing issues: managing participant confidentiality”

UNIVERSITY OF OXFORD

A few thoughts about today’s projects

- *Dig That Lick*
- Based on symbolic mark-up of audio
- Essential to sustain the basic audio
- Code sustainability to reproduce the analyses — tricky
- Like speech: nice to combine signals with mark-up

Sustainability: gold standard

Even better: rock standard (cf. MTAAC)

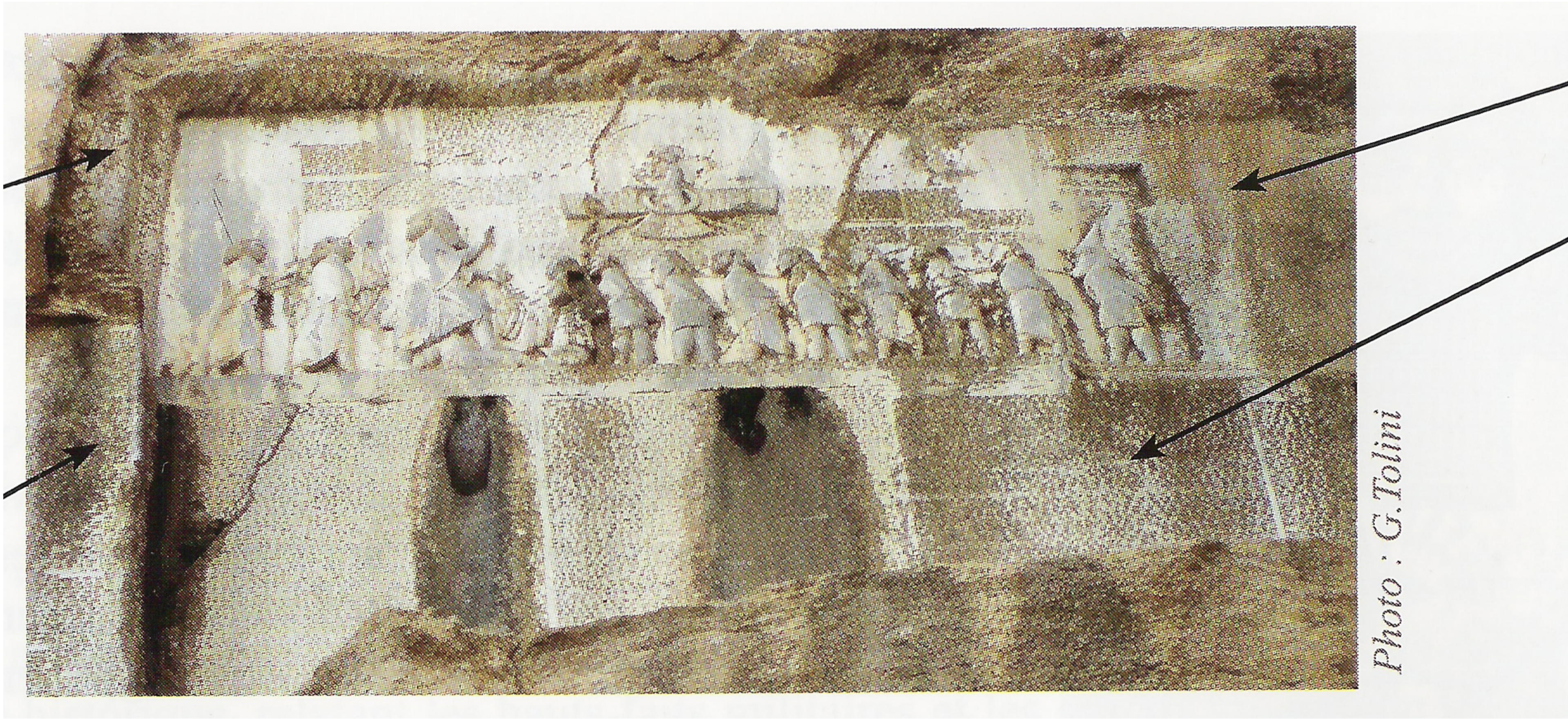
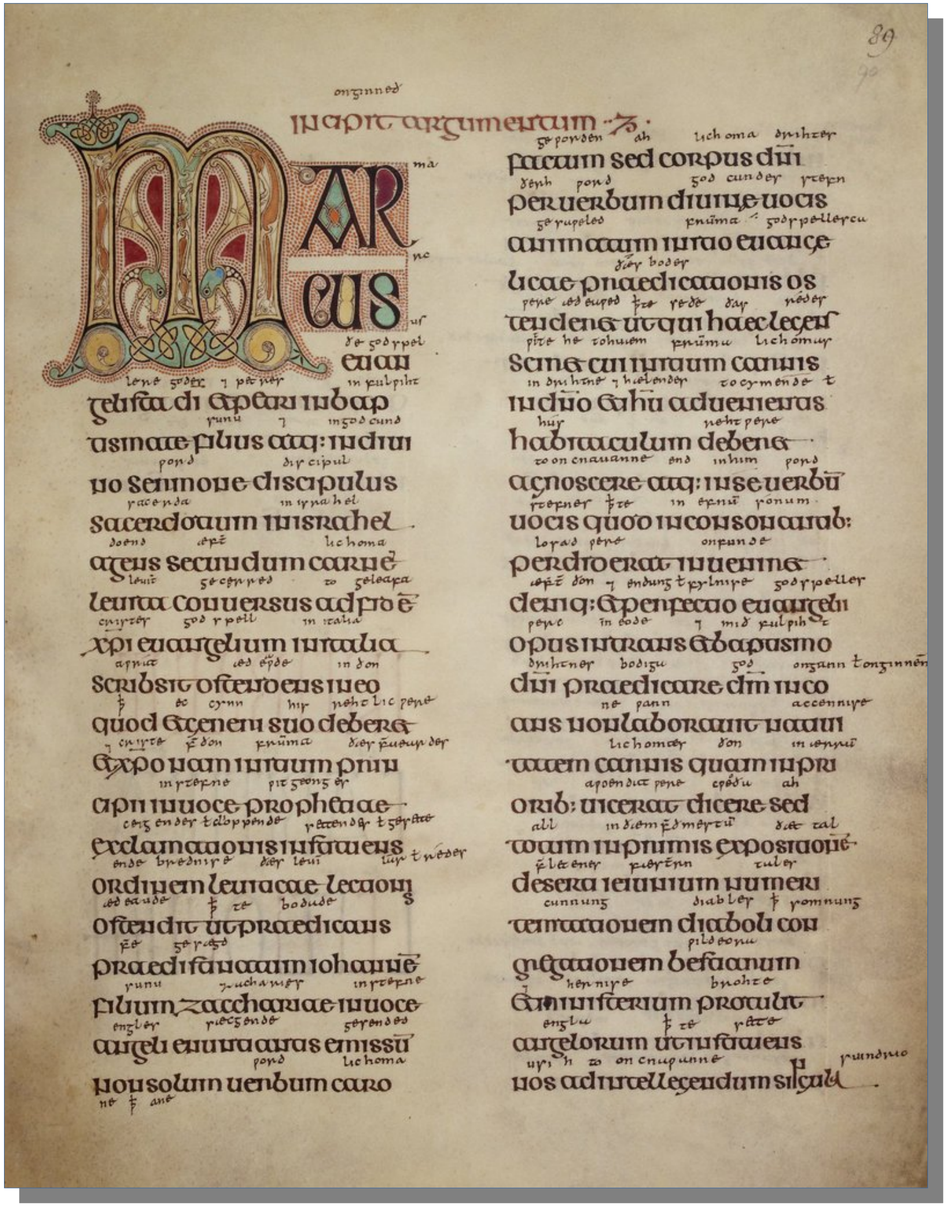


Photo : G. Tolini

Sustainability: also good



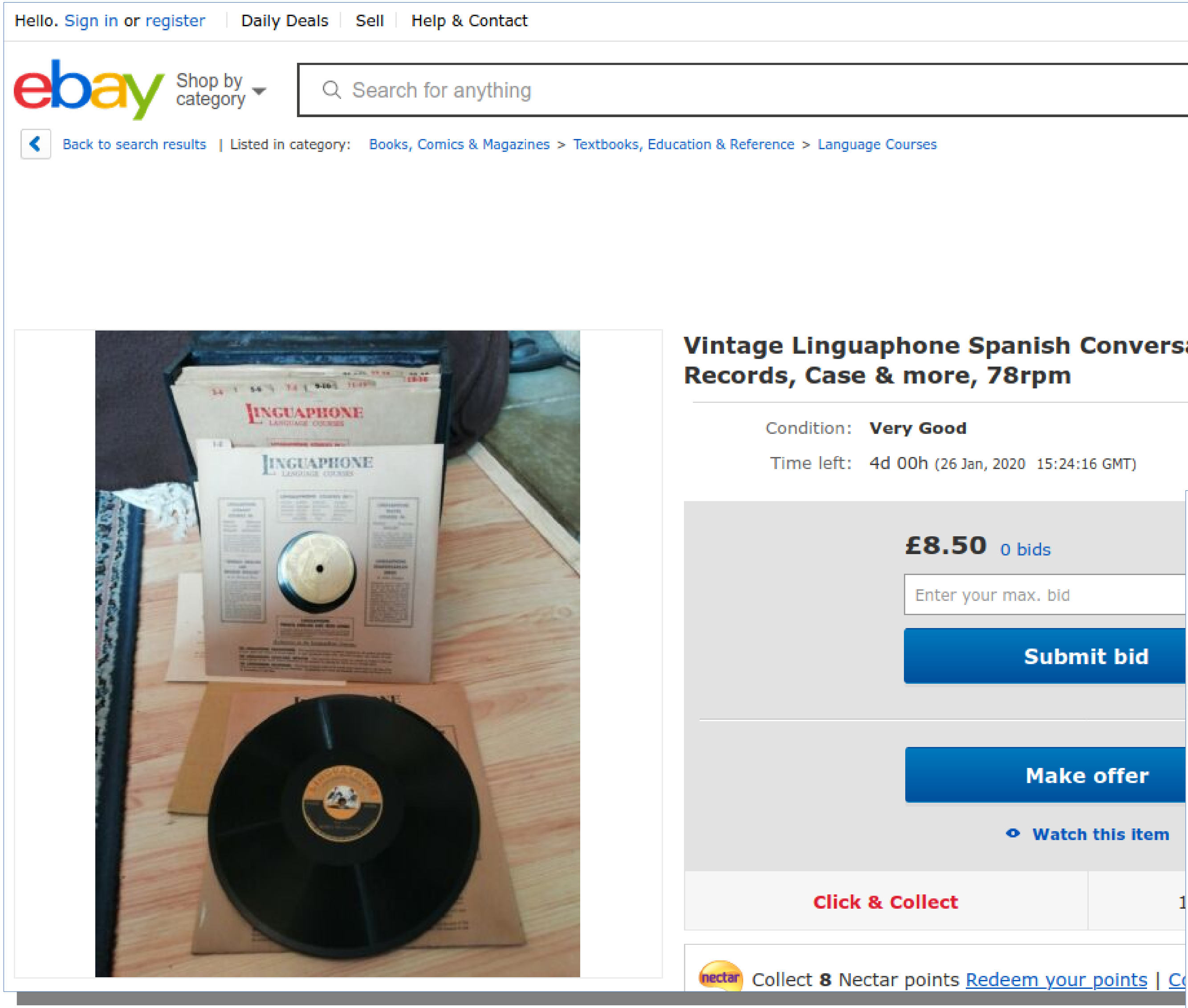
89
90

INCIPIT ARGUMENTUM

MARCUS euan
gelista di electus et petri in bap
tismate filius atque in diui
no sermone discipulus
sacerdotium in israhel
agens secundum carnē
leuita conuersus ad fidē
xpi euangelium in italia
scribsit ostendens in eo
quod et generi suo deberet
et xpo nam initium prin
cipii in uoce propheticae
exclamationis instituens
ordinem leuiticae lectionis
ostendit ut praedicans
praedistinatum iohannē
filium zachariae in uoce
angeli enuntiantis emissū
non solum uerbum caro

factum sed corpus dn̄i
per uerbum diuinae uocis
animatum initio euange
licae praedicationis os
tenderet ut qui haec legens
sciret cui initium carnis
in dn̄o et dei aduenientis
habitaculum deberet
agnoscere atque in se uerbū
uocis quod in consonantib:
perdiderat inueniret
denique et perfecti euangelii
opus intrans et a baptismo
dn̄i praedicare dm̄ inco
ans non laborauit natiui
tatem carnis quam in pri
oribus uiderat dicere sed
totum in primis expositione
deserti ieiunium numeri
temptationem diaboli con
gregationem bestiarum
et ministerium protulit
angelorum ut instituens
nos ad intellegendum singula

UNIVERSITY OF OXFORD



Hello. Sign in or register | Daily Deals | Sell | Help & Contact

< Back to search results | Listed in category: Books, Comics & Magazines > Textbooks, Education & Reference > Language Courses

Vintage Linguaphone Spanish Conversational Course 16 Records, Case & more, 78rpm

Condition: **Very Good**

Time left: 4d 00h (26 Jan, 2020 15:24:16 GMT)

£8.50 0 bids

Collect 8 Nectar points Redeem your points | Co

Speech recordings

- Fortunately, many archives and libraries hold collections of lots of audio material
- Unfortunately, a lot of it is in formats that are largely/often inaccessible
- Fortunately, a lot of it is being digitized and can be copied perfectly
- Unfortunately, too many language researchers and archivists are unduly possessive about their collections and inhibit access to it by others

My sustainability recommendations

- Put your data directly on the internet (not mediated by app/login/search tool etc)

- Use some standard syntax for data objects

- e.g. bnc.phon.ox.ac.uk/data/021A-C0897X0229XX-ABZZP0.wav?t=1560,1582

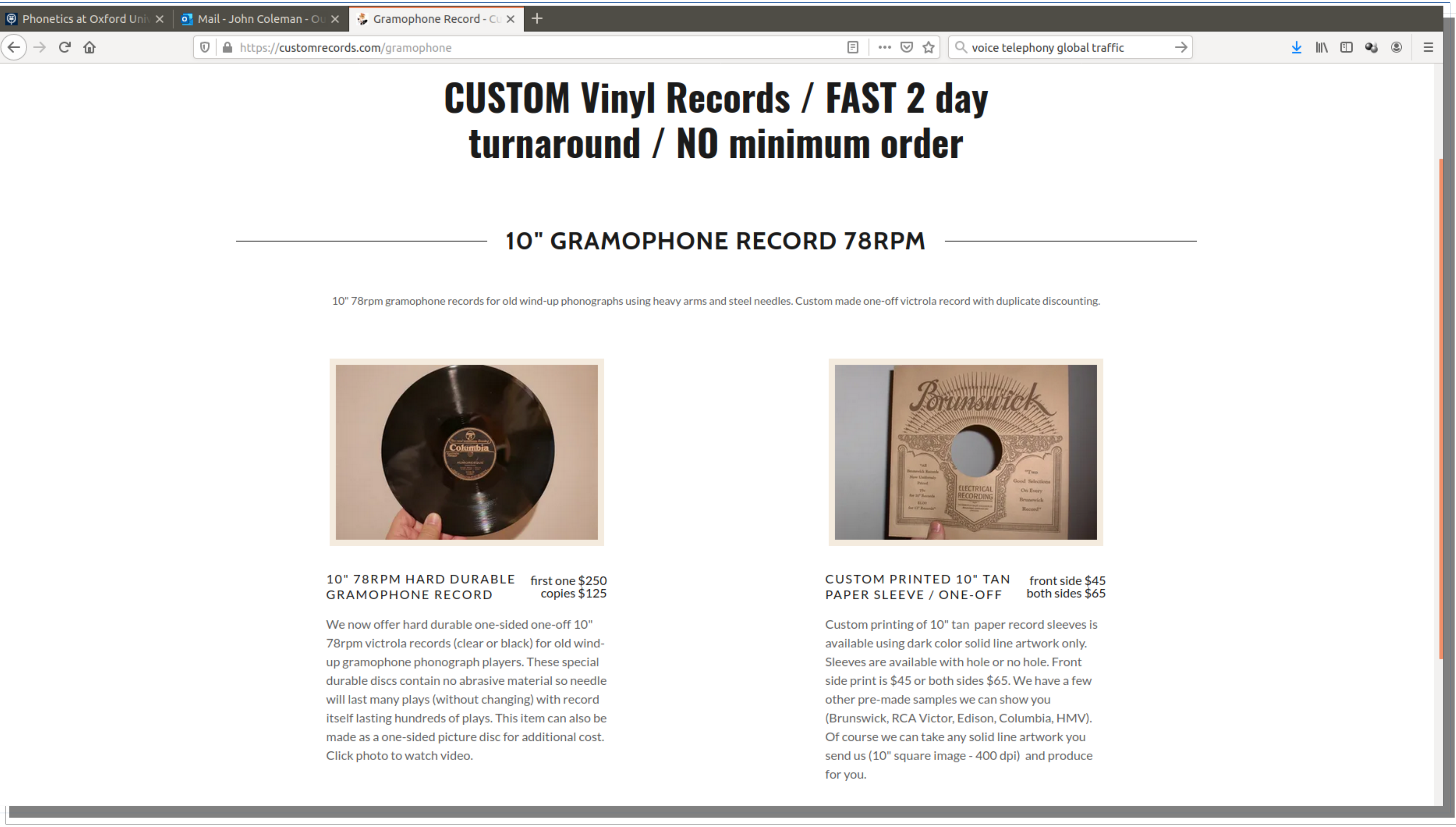
- Don’t change URL’s/URI’s if you reorganise your site

My sustainability recommendations

- Allow/encourage free copying

- Don’t *assume* that participants want *secrecy*; consent for openness is highly desirable for scientific probity as well as sustainability. *Anonymity* can allow for openness

- Engrave backups on some suitable solid material



CUSTOM Vinyl Records / FAST 2 day turnaround / NO minimum order

10" GRAMOPHONE RECORD 78RPM

10" 78rpm gramophone records for old wind-up phonographs using heavy arms and steel needles. Custom made one-off victrola record with duplicate discounting.

10" 78RPM HARD DURABLE GRAMOPHONE RECORD

first one $250
copies $125

We now offer hard durable one-sided one-off 10" 78rpm victrola records (clear or black) for old wind-up gramophone phonograph players. These special durable discs contain no abrasive material so needle will last many plays (without changing) with record itself lasting hundreds of plays. This item can also be made as a one-sided picture disc for additional cost. Click photo to watch video.

CUSTOM PRINTED 10" TAN PAPER SLEEVE / ONE-OFF

front side $45
both sides $65

Custom printing of 10" tan paper record sleeves is available using dark color solid line artwork only. Sleeves are available with hole or no hole. Front side print is $45 or both sides $65. We have a few other pre-made samples we can show you (Brunswick, RCA Victor, Edison, Columbia, HMV). Of course we can take any solid line artwork you send us (10" square image - 400 dpi) and produce for you.

For info only: not an advertisement

UNIVERSITY OF OXFORD