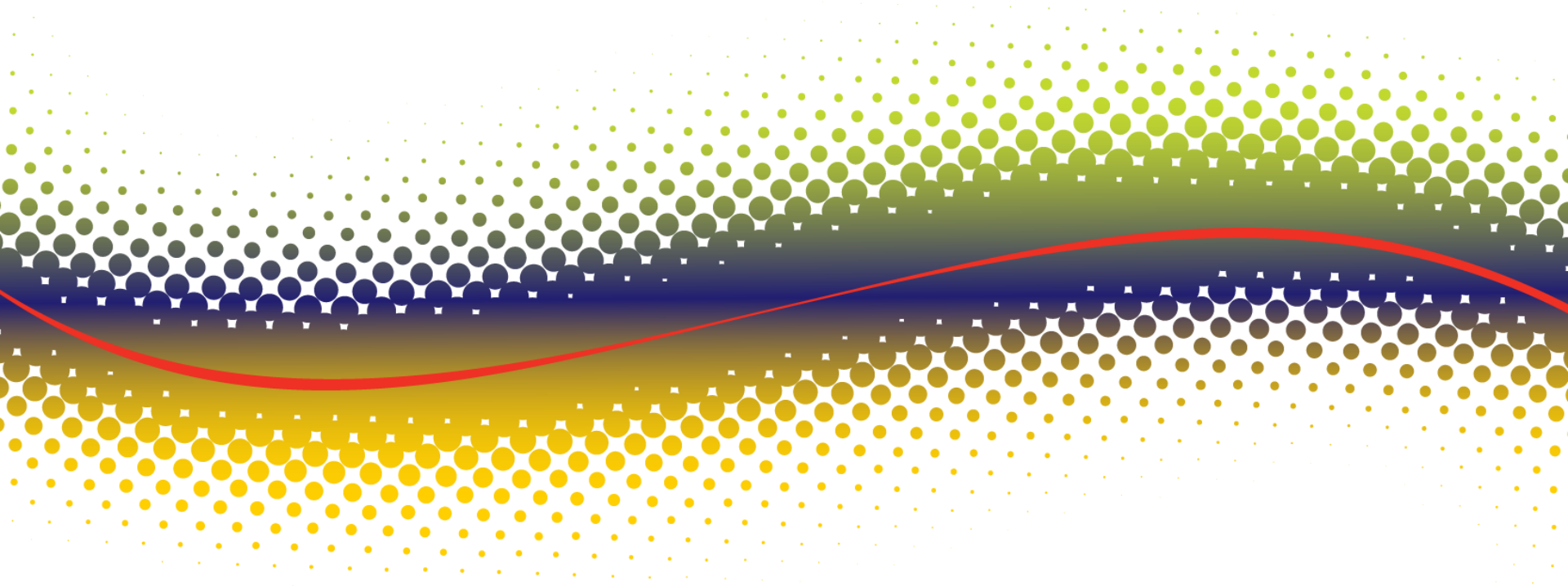
Linguistic Data Consortium

Language Resources of and in the Americas: Challenges and Opportunities

Christopher Cieri, James Fiumara



Linguistic Data Consortium

Why Are We Here

- Support for Networking in US < Europe
  - NetDC, Digging into Data experiences
  - FlareNet
  - CLARIN
- COCOSDA
  - Early: LTTS Server
  - Current: biennial with regional, topical reports
  - OCOCOSDA: annual, contribution on languages of Asia
  - Shortcomings: speech only, isolates data creators from users
- Opportunities
  - Increasing participation from ‘the Americas’
  - Penn Global Initiatives
    - China, Latin America, most recently India
- Can we organize to better collaborate

Why Are You Here

- Some participants commented that you were:
  - not a data center
  - not working specifically on languages of the Americas
- But You Represent
  - Data creators
  - Institutional Archives
  - 'Regional' Data centers
  - Search Providers
  - Networking Services
- Working either on
  - Languages of the Americas
  - Language in the Americas
- Effect of that context is unique (or at least different from Europe) in terms
  - linguistic, demographic, economic, legal/regulatory, cultural
- Frankly most of you represent multiples of these

Workshop Goals

- Meet to sketch functions/approaches
- Identify Major Challenges
- Can the challenges faced by any one group be met by any other?
- Identify challenges in common
- Evaluate according to importance, urgency
- Do we have enough in common to justify greater networking, collaboration?
- If so, what would that look like?

Linguistic Data Consortium

LDC Functions/Issues

- Publications
  - first function
  - generally successful
  - QC all publications upon deposit, work with contributor to fix, rarely reject
  - required normalizing processes so that regardless of delays in submission, publish >= 30 data sets per year
  - create queue
  - plan publications annually
    - vary topics to match member interests
  - respond to well formed requests to vary schedule

LDC Functions/Issues

- Archiving/Curation
  - required modernizing storage
    - Service Center
    - regular replacement
    - review of all data by type for static/dynamic storage, implementing IRODS
    - data integrity
    - disaster recovery
  - issue with benchmark corpora
  - QC upon receipt
  - mechanism for patches
  - re-issues to match modern uses

LDC Functions/Issues

- Membership
  - developing an approach to marketing & communications
  - revised our distribution to have an e-commerce feel
  - moved from CD -> DVD -> BD to HD + USB + cloud based distribution
    - local and via cloud providers
  - biennial surveys
    - members nearly always satisfied to very happy
    - complaints generally reflect communications breakdowns
      - misunderstanding of LDC model
      - cost
  - data cost
    - generally well understood
    - free/low cost corpora
    - data scholarships

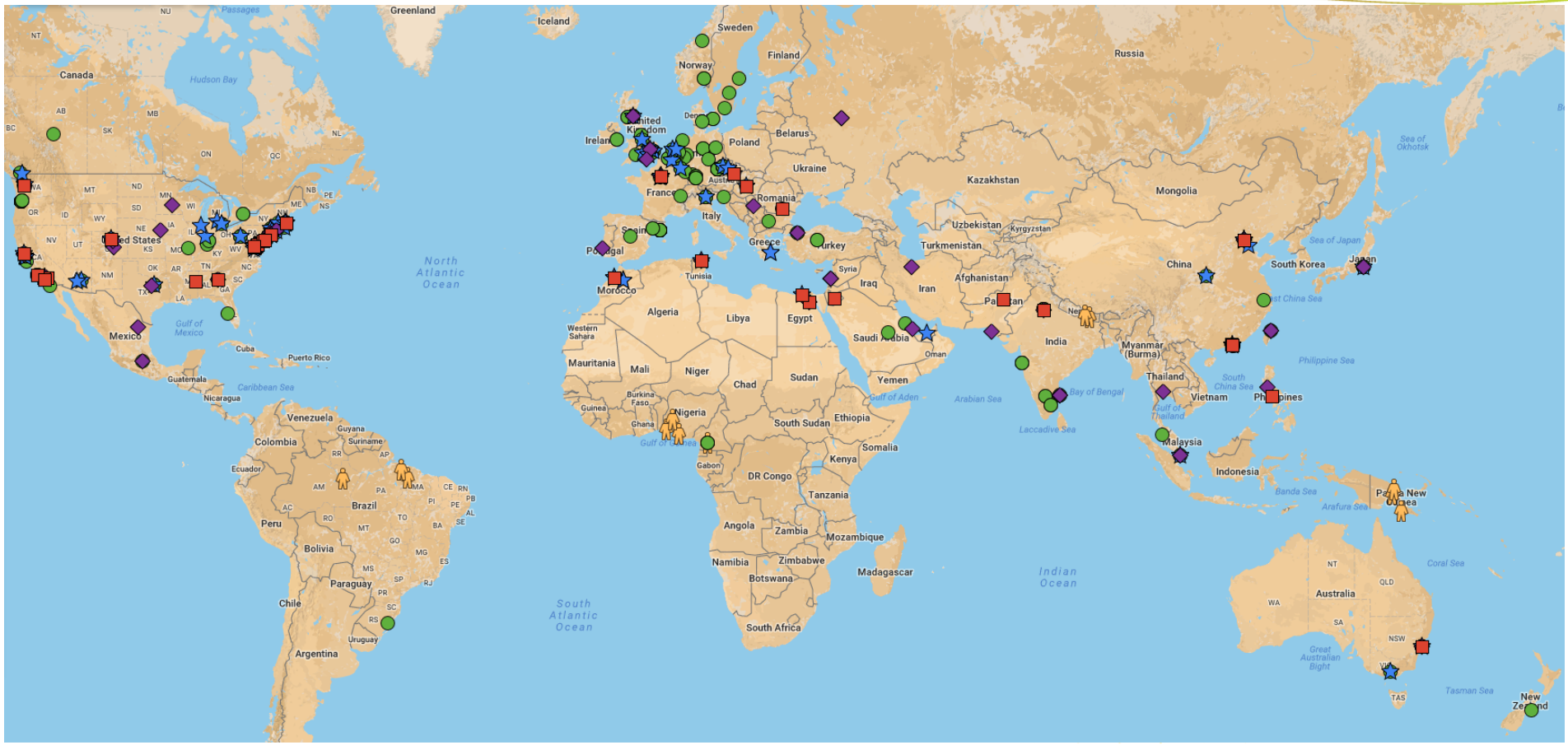
Linguistic Data Consortium

LDC Functions/Issues

- Collect & Annotate
  - already on a large scale and good diversity
  - however major challenges remain
    - diversity (Zulu)
    - scale
    - skills/training

Linguistic Data Consortium

Background

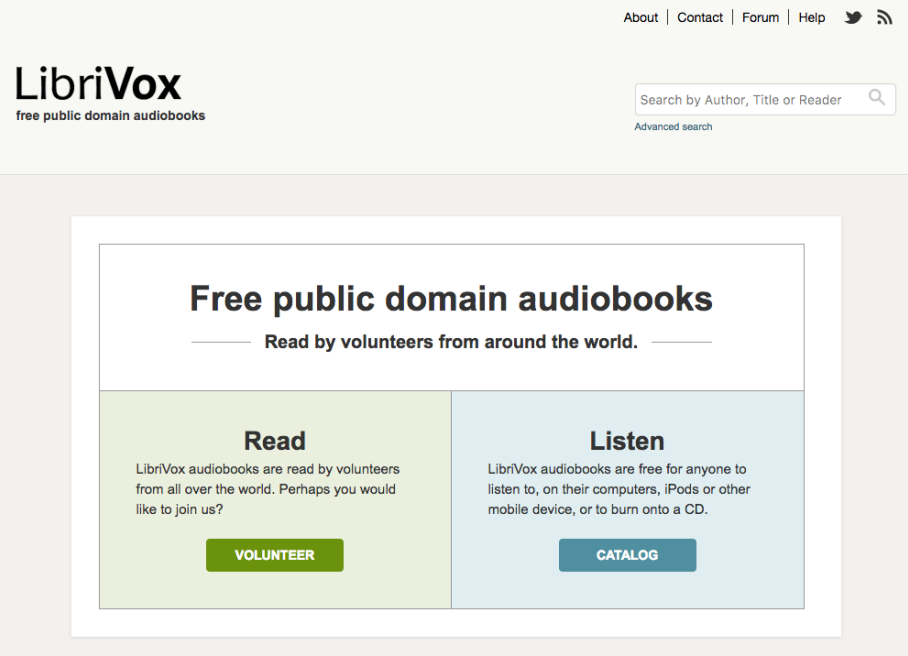


Some LDC data sources including subcontractors and vendors (red squares), corpora (green circles), media providers (purple diamonds), LDC staff collections (gold people), research collaborators (blue stars). Many markers represent multiple collaborators; some markers partially obscured by others

Features of NIEUW Approach

- always available to anybody, on any device
- though some activities require additional skills, data
- connection social networking sites for recruitment, data
- incentives: information, entertainment, self-expression, socializing, developing skills, demonstrating competence, competition, status, prestige, recognition, access to (HLT) services, contributing to a greater cause or good
- data prioritization replaces data selection
- X-sourcing: tasks selected, engineered to match workforce, incentives
  - crowd, gamers, citizen scientists, language professionals
- automated worker training and assessment
- algorithms for task assignment, finding experts, modeling variation
- integration with current LDC operations, including crowd-sourcing

LibriVox



- LibriVox
  - "*free public domain audiobooks*"
  - recruit, train, organize volunteers who
  - record readings of literary works out of copyright in US

- January 1, 2017
  - 53,430 hours of readings in English
  - 8,689 hours in 36 other languages
  - 1 recorded hour requires 2 hours of reading time + 2-4 hours editing > 248,476

- At industry standard rates, $500 per finished hour of audio, cost to fund LibriVox would exceed $31,000,000.

- Incentives
  - belief in LibriVox mission
  - interest in content
  - promoting specific work, author, genre
  - challenge of acting without visuals
  - opportunity to develop and apply skills; spin-offs to Iambik and Audible
  - work with like-minded collaborators in multiple sub-communities
  - enlightened self-interest

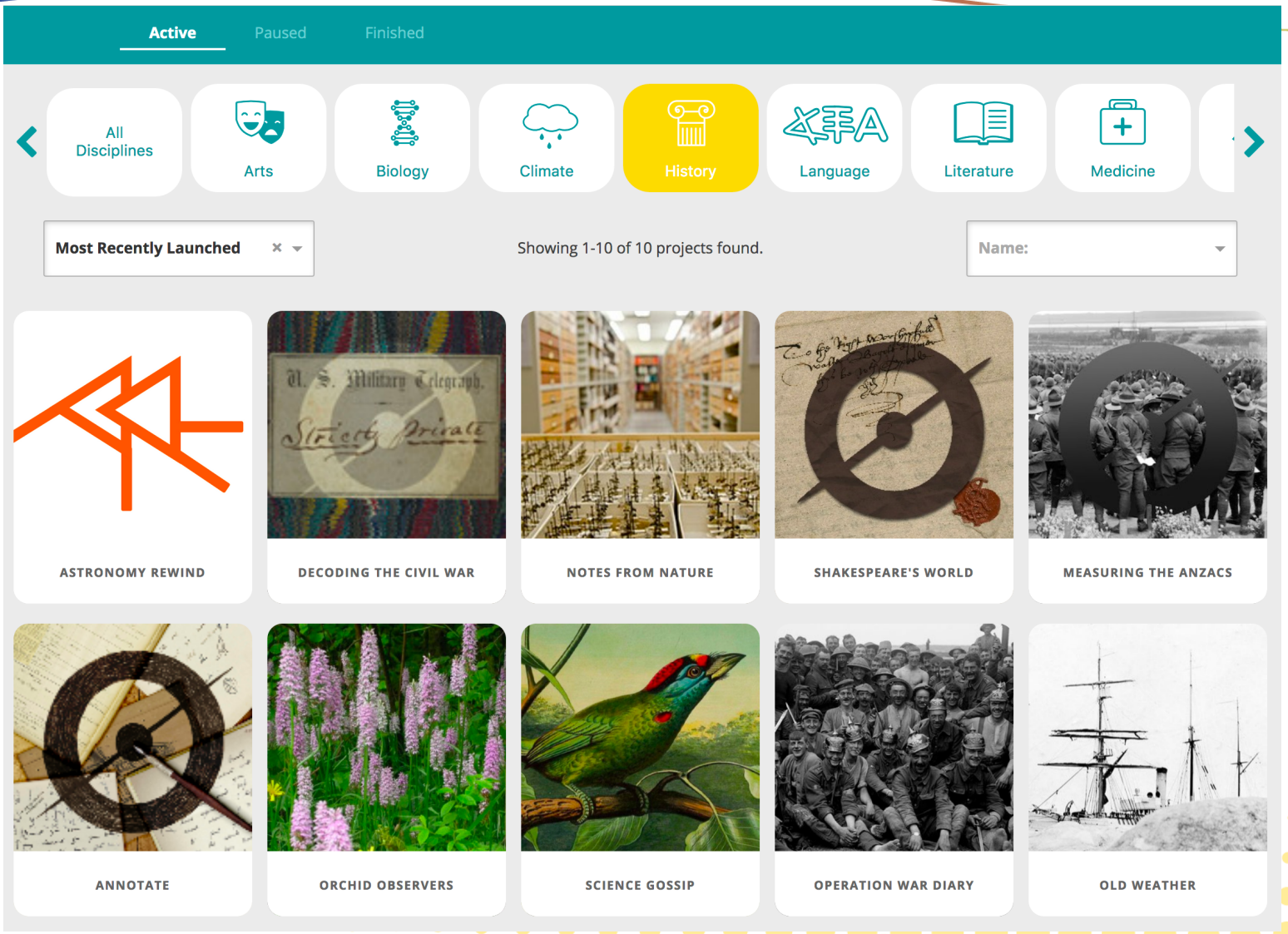
Great Language Game

- Great Language Game
  - players hear brief audio clips in ~80 languages
  - select correct language from multiple choice
  - 3 lives
    - correct = 50 points
    - incorrect = lost life
  - # choices increases during game
  - Ethnologue entries for languages missed

- History
  - 2013 appeared
  - 2014 public data release, < 1 year later
  - >16,000,000 judgments
  - > all NIST LRE judgments ever
  - now >38,000,000
  - stats disappeared
- Players include: Central & South America, Sub-Saharan Africa, Central Asia, Oceania.
- Incentives
  - information
  - entertainment
  - competition
  - status
- Judgments suboptimal for LRE
  - too few clips
  - too many judgements/clip
  - all answers known

Linguistic Data Consortium

Citizen Science: Zooniverse



Components of NIEUW

- Data, Existing Annotations
  - existing at LDC currently being reindexed, ongoing collection
  - new contributions, collection from social network, smart devices
  - NIEUW collection activities
- Annotation Tools
  - WebAnn adjusted for: less expert task creators, citizen scientists, portability, mobile users, 'reporting' to contributors
  - language game building toolkit from Essex
- Portals
  - language professionals
  - citizen scientists
  - language game players
- Activities
  - subset created as proof-of-concept in infrastructure project
  - others added to respond to or anticipate demand
- Partners
  - contribute activities
  - host their own instances

Linguistic Data Consortium

Games: Great Language Game

LINGO BOINGO

World Language Games

All English French

Jeux de mots

French

Lexical and semantic games with a purpose in French.

Phrase Detectives

English

Compete against other detectives by identifying the relationships between words and phrases in a variety of texts including literature, history, travel, entertainment and science. Earn top scores!

Tile Attack

English

Go head-to-head against another player competing to identify the noun phrases of a text.

Zombilingo

French

Identify syntactical dependencies, collect brains and eat them! This language game is fun for both fans of grammar and zombies.

Relevant Activities

- Basic Collection/Annotation Tasks
  - oral history, my best X ever
  - web video accessibility via transcription, translation
  - massively parallel multilingual dictionary, multimedia language atlas
  - transcription, LID, HASR, ERE, NP chunking, head hunting
- Annotating Situation Frames
  - as Zooniverse already does in ANZACS date, nationallity, height, weight, for a kind of SF
- Emergency Social Media
  - collect multilingual social media posts from previous natural disasters
  - create new posts via scenarios
  - translate posts into and from English
- Opinion Spectrum
  - for 'news topic' find stories, segments representing range of opinions: pro -> con
- Call BS or The Smell Test
  - for given statement, express confidence, provide independent evidence
- Also any stable annotation task, divided into atomic tasks.

Work Outline

- NSF CRI planning grant, NSF CRI NEW proposal awarded
- presentations, discussions
  - NSF, DARPA, IARPA, DOD, CASL, AFRL
  - MIT-LL, CLARIN, SADiLaR
  - LREC, NSF workshops on Novel Incentives and Workflows
  - program committee, etc. for LREC, Games4NLP, HCOMP
- Development Partners
  - Uni Essex, Sorbonne, enetCollect
- New Data Contributors
  - Labov, Drager, D'Arcy, Feigin, Yaeger-Dror
- WebAnn ported
- First game nearly finished