Linguistic Resources for the 2015 TAC KBP Cold Start and Tri-Lingual Entity Discovery & Linking Evaluations

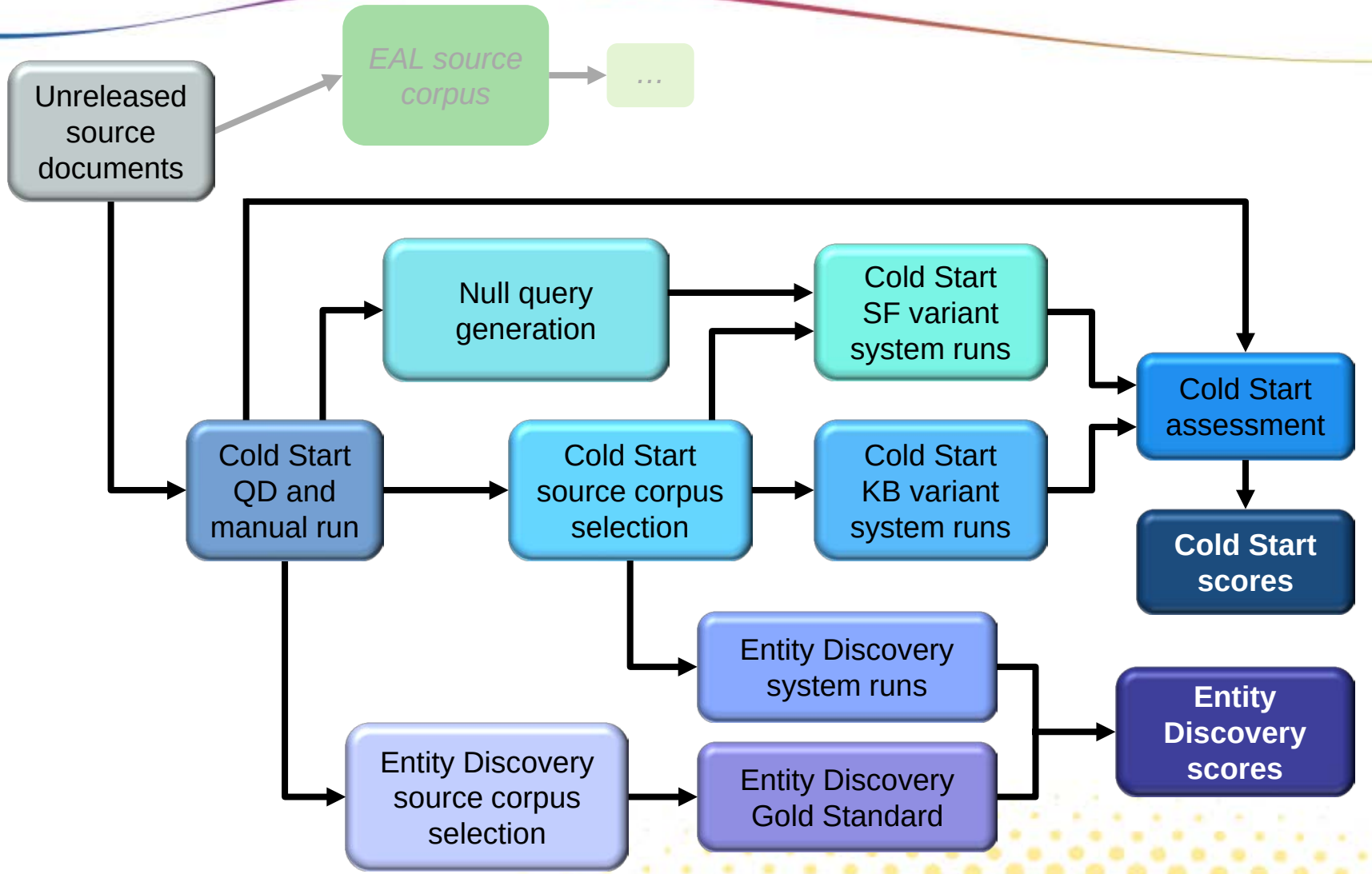
Joe Ellis (presenter), Jeremy Getman, Stephanie Strassel

Linguistic Data Consortium
University of Pennsylvania, USA

Cold Start

- For data development purposes, Cold Start is a question answering task
  - Since 2012, LDC has approached Cold Start from the ‘Slot Filler Variation’ perspective
  - We’ve never previously had to concern ourselves much with the KB construction side of the task.
- However, query requirements changed significantly for 2015
  - More on this later…

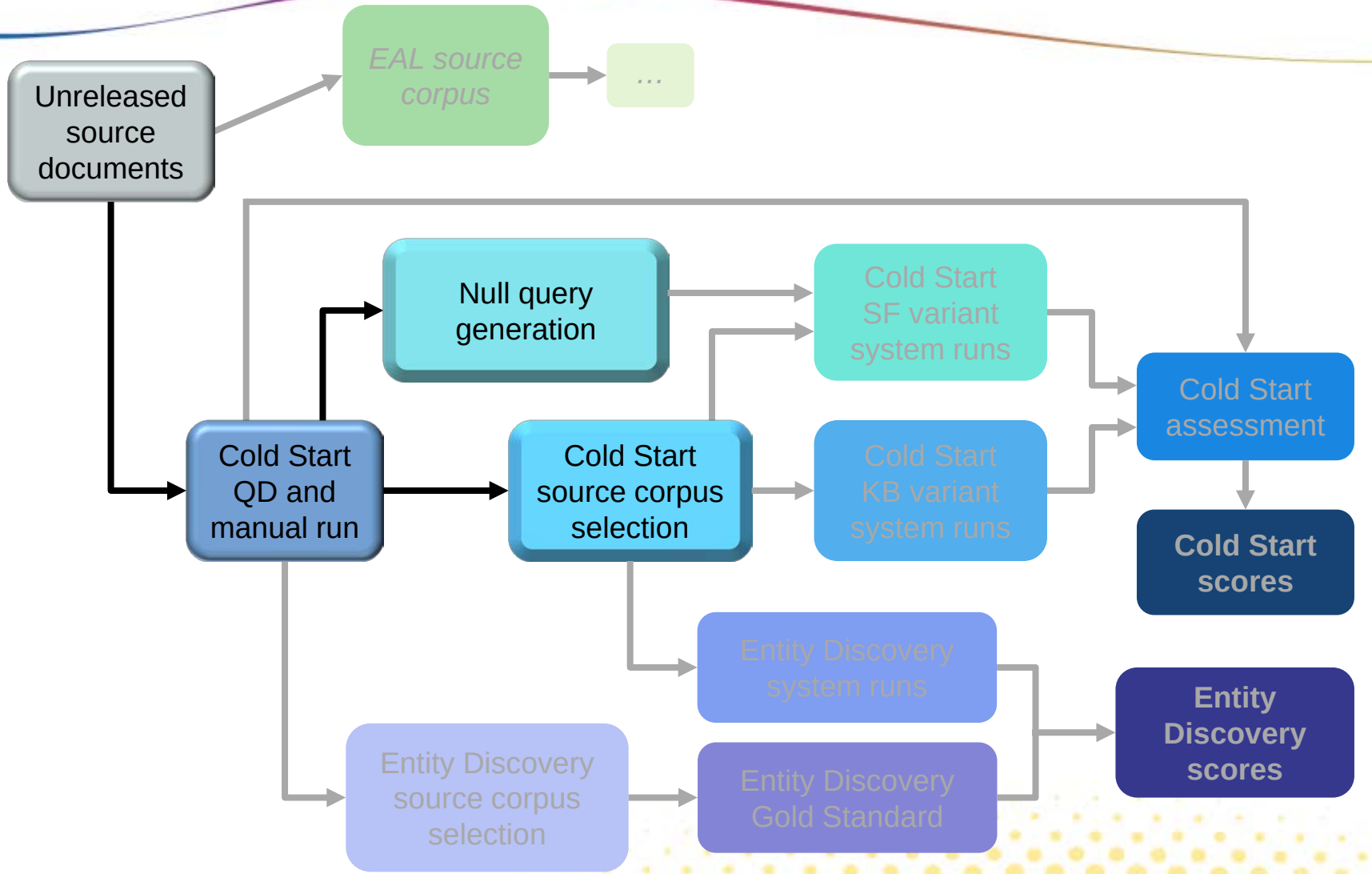
Cold Start Data Pipeline



Cold Start: Source Document Pools

- Three pools of unexposed documents
  - 2013 New York Times articles
    - ~57,000 documents
  - 2013 Xinhua articles
    - ~190,000 documents
  - Multi-post Discussion Forum threads (MPDFs)
    - Truncated discussion forum threads
    - Over 1 million MPDFs
- Annotators searched document pools to develop queries and the manual run
- Additional documents for the final source corpus also selected from these pools

Cold Start QD & Manual Run



Cold Start QD & Manual Run

Lance Barrett, 23, of London, KY, was charged with first-degree attempted burglary, theft of a firearm, and carrying a concealed weapon.

Lesa Bailey, 44, of London, KY, was charged with criminal conspiracy to make meth, unlawful possession of meth precursors and possession of a controlled substance.

London – *gpe:residents_of_city* – *per:charges*

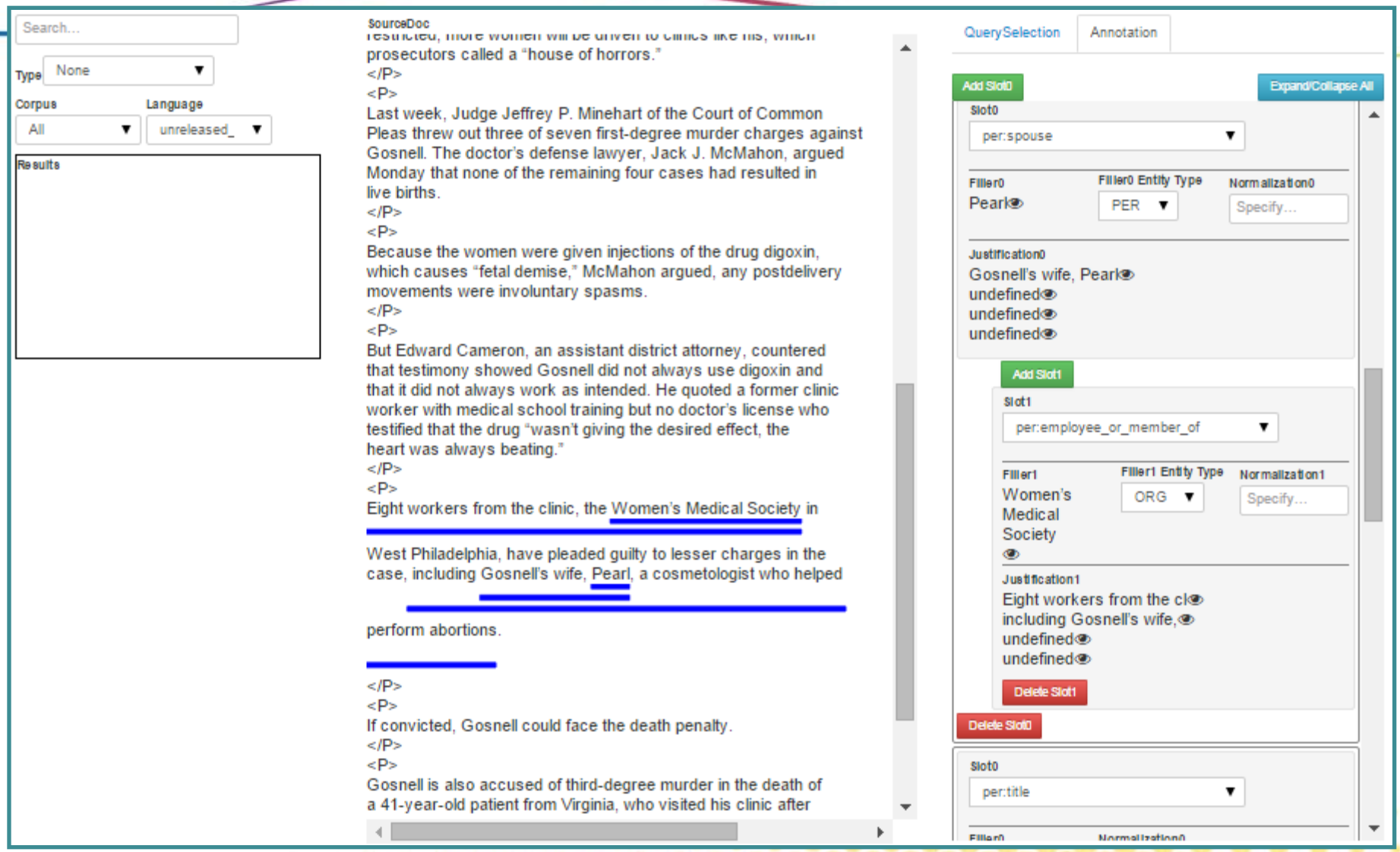
- Lance Barrett
  - first-degree attempted burglary
  - theft of a firearm
  - carrying a concealed weapon
- Lesa Bailey
  - criminal conspiracy to make meth
  - unlawful possession of meth precursors
  - possession of a controlled substance

- Chains of entities connected by KBP slots
  - Cold Start queries comprised of
    - Entity – *Slot 0* – *Slot 1*
- Cold Start annotation & query development concurrent
  - Annotators attempt to balance
    - Targeted number of annotations
    - Query variety (entity type, slot type, genre, etc.)
  - Annotation not exhaustive – some slots are more productive than others

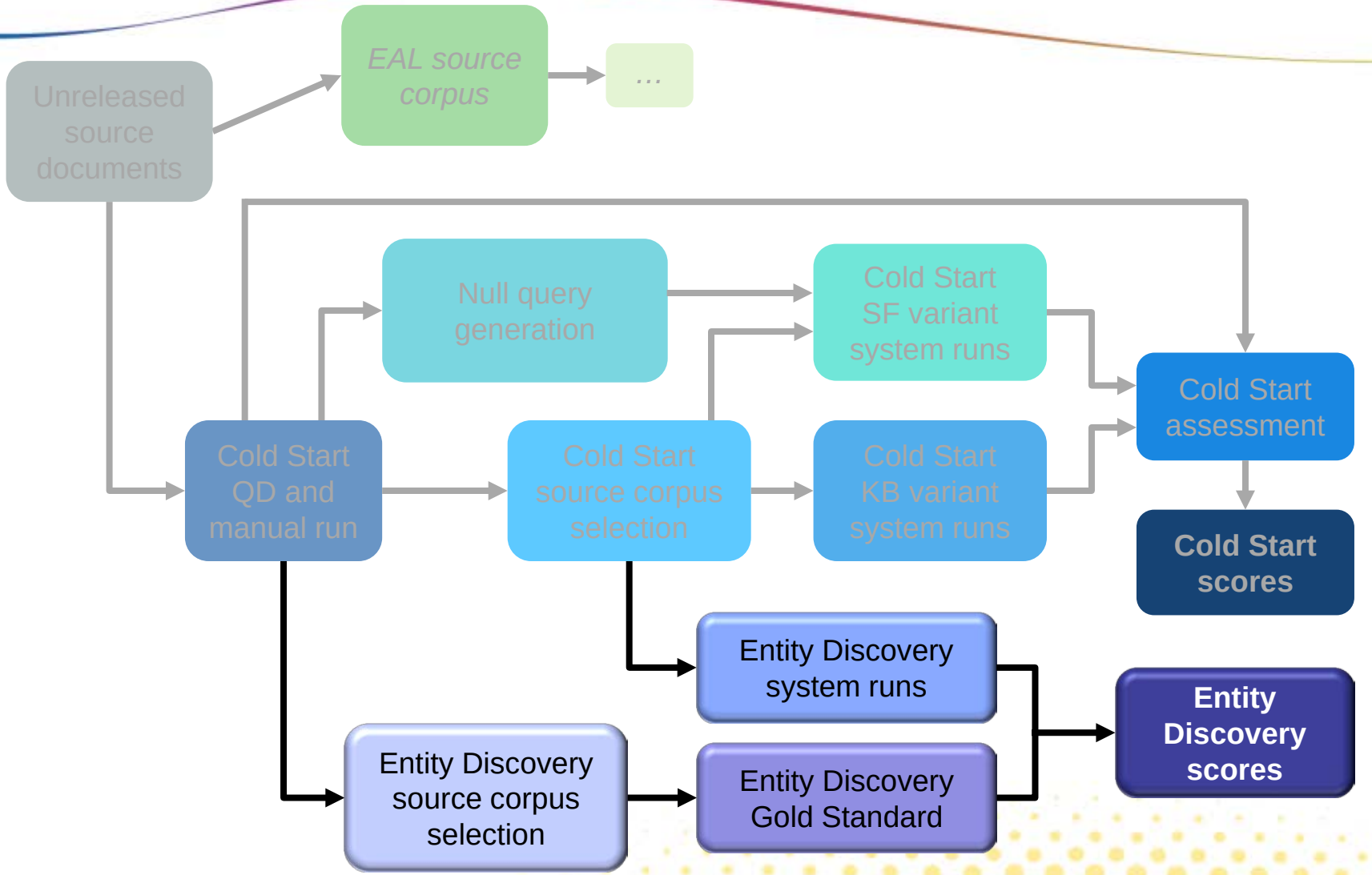
Cold Start: Query Development Changes

- Changes to query requirements compared to 2014 data
  - High degree of overlapping Entry Point Entities (EPEs) across queries
    - 2-5 mentions from different sources
    - Ambiguous whenever possible
  - Null queries
    - Auto-generated for rapid production but not guaranteed to have no valid responses
  - Changes made primarily to support Slot Filling subsumption and to ensure challenge for Entity Discovery

LDC's Cold Start GUI



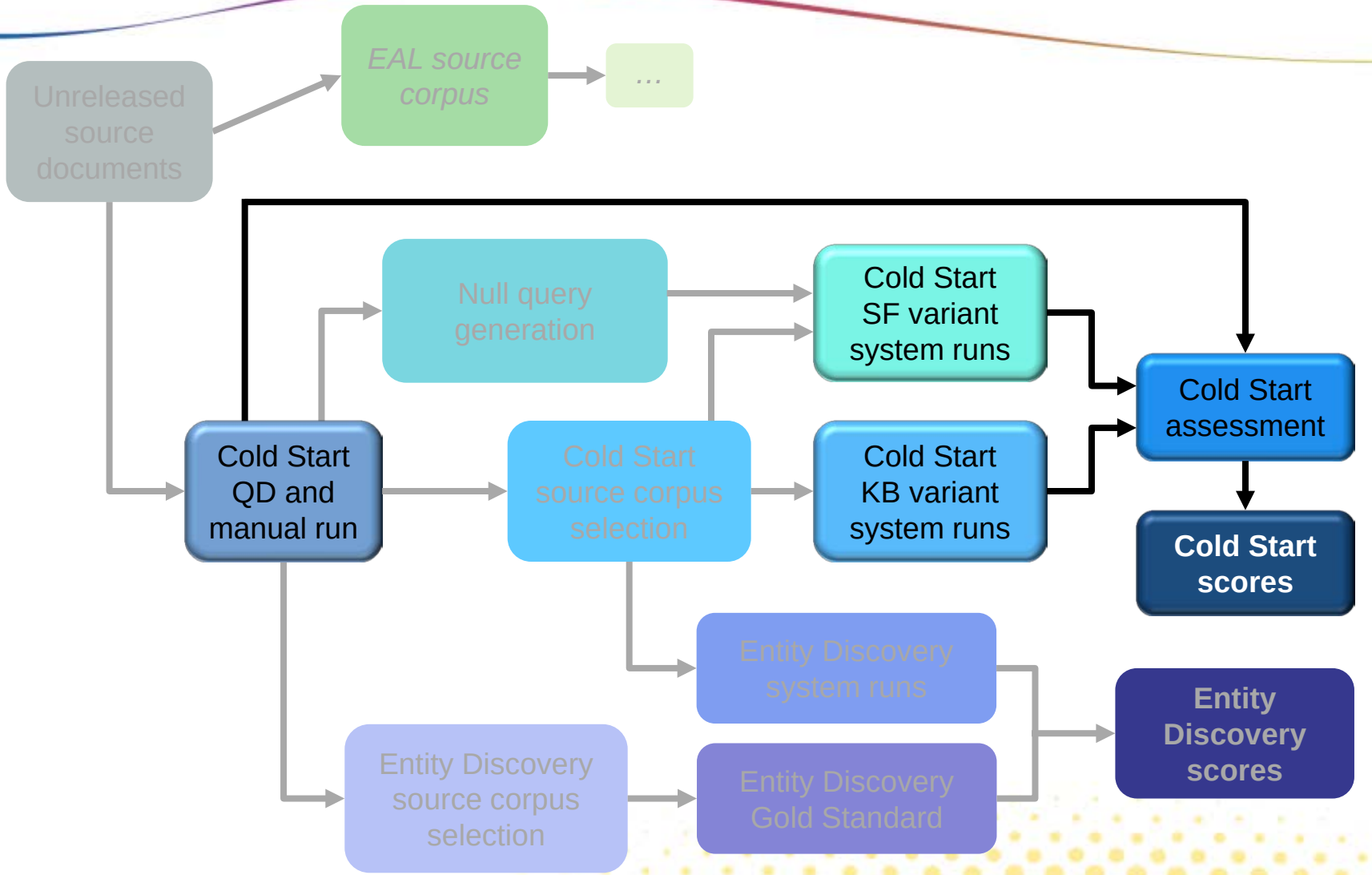
Cold Start: Entity Discovery



Cold Start: Entity Discovery

- Identifying and clustering all valid entity types in the Cold Start corpus
  - Effectively, simplified Entity Discovery & Linking
    - One language, less entity types, one mention type
- Gold Standard development
  - As in ED&L, Cold Start – Entity Discovery submissions were scored against LDC's gold standard mentions
  - 200 document subset of Cold Start source corpus
  - High degree of overlap with Cold Start queries and manual run
  - Entities mentioned in multiple documents, some with ambiguous mentions

Cold Start: Assessment



Cold Start Assessment

- NIST pools results and sends to LDC
- Assessment performed in batches
  - hop-0 and hop-1 responses assessed for subset of queries
  - Queries to be assessed were selected and batched by NIST
- Assessment continues in batches until resources exhausted

Cold Start Assessment

- Assess validity of fillers & justification from humans & systems
  - Filler
    - Correct – meets the slot requirements and supported in document
    - Wrong – doesn't meet slot requirements and/or not supported in doc
    - Inexact – otherwise correct, but is incomplete, includes extraneous text, or is not the most informative string in the document
  - Predicate
    - Correct – provides all information necessary to link the query entity to the filler by the chosen slot
    - Wrong – does not provide any of the necessary information
    - Inexact-Short – provides some, but not all, of the necessary information
    - Inexact-Long – otherwise correct, but includes extraneous text
- Correct and Inexact responses clustered together

Cold Start: Assessment Results

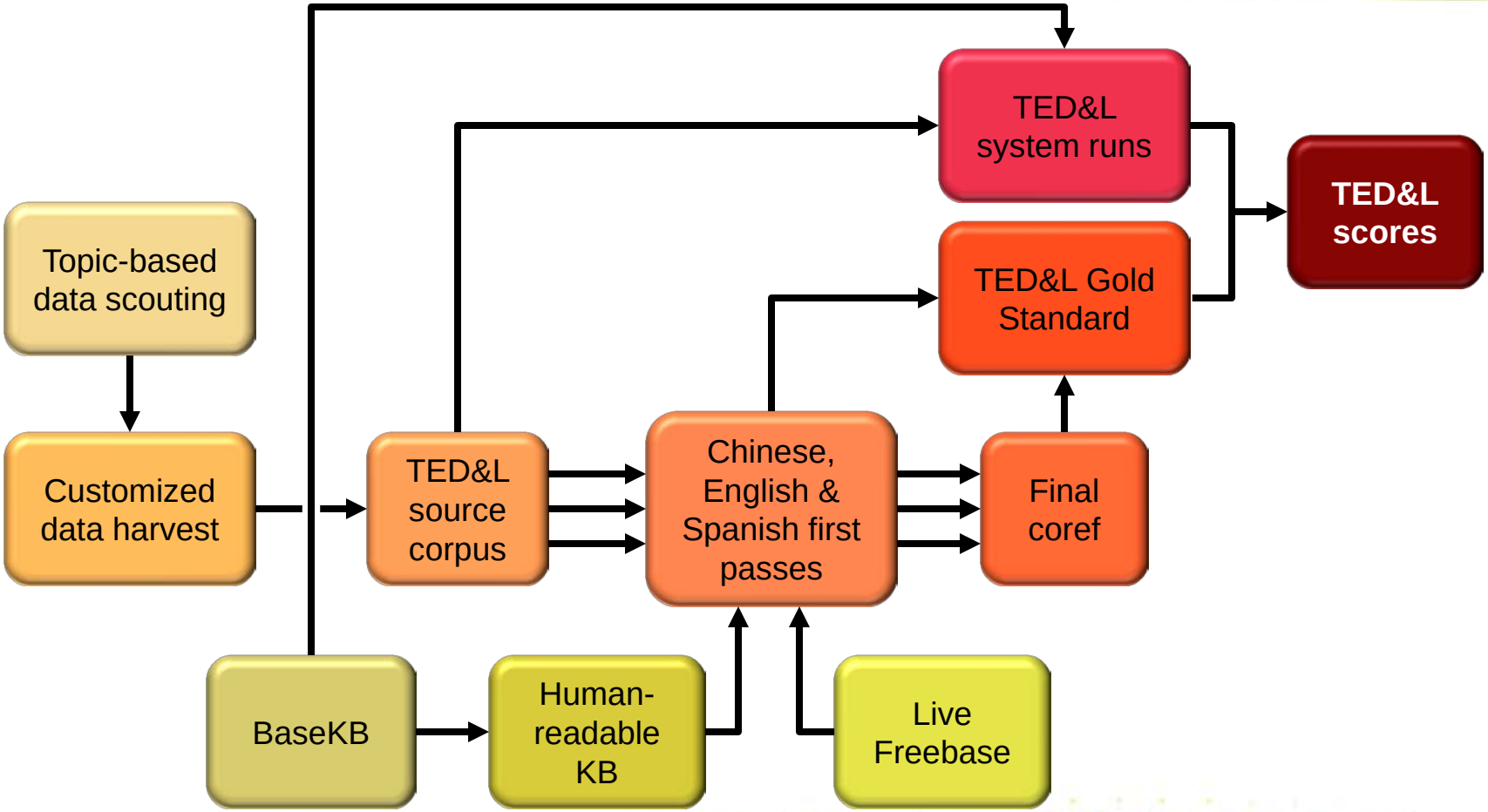
	Total	Newswire	MPDF
Responses	30,654	15,948	14,706
Correct	26.7%	29.7%	23.5%
Wrong	68.8%	65.2%	72.8%
Inexact	4.5%	5.1%	3.7%

Cold Start: Manual Run Results

Track	Precision	Recall	F1
2014 Cold Start	91%	46%	62%
2015 Cold Start	81%	19%	31%

- New approach allowed for better tracking of query requirements, but may have further reduced focus on manual run
  - Focus given to competing query requirements
  - Annotators less exacting when selecting fillers
- Inexact responses included in scoring
- More queries
  - 1,327 productive queries (not including hop-1 portions)
  - 750 queries for Cold Start, Sentiment SF and Regular SF combined in 2014

TED&L Data Pipeline



TED&L: Changes from 2014 ED&L

- New knowledge base
- New source data requirements
- New annotation requirements
  - Monolingual to tri-lingual
  - New entity types
    - FAC & LOC
  - New mention type
    - NOM

TED&L: New Knowledge Base

- The old KB (2008 Wikipedia snapshot) made task too artificial
- Distributed via two releases
  - BaseKB (basekb.com)
    - FreeBase converted into RDF
  - Algorithm for creating for KB entries
    - Describes process by which triples were collected into pages for annotators to review

BaseKB	2008 Wikipedia Snapshot
As a triple store, systems can interact with the KB as a graph	Only available as a collection of entries
Over a billion facts about more than 40 million subjects	818K entries

TED&L: Source Data

- Requirements
  - 500 documents
  - Cross-lingual, cross-document entity mentions
  - Multiple, varied, recent sources
- Challenges
  - Unusual approach for harvesting/processing
    - Usual approach is larger volumes from fewer sources
    - Additional effort required
  - Managing Intellectual Property Rights issues
    - Ensuring LDC has the right to annotate and redistribute collected data
    - 100s of sources required new approaches new approaches
      - Data distribution formats

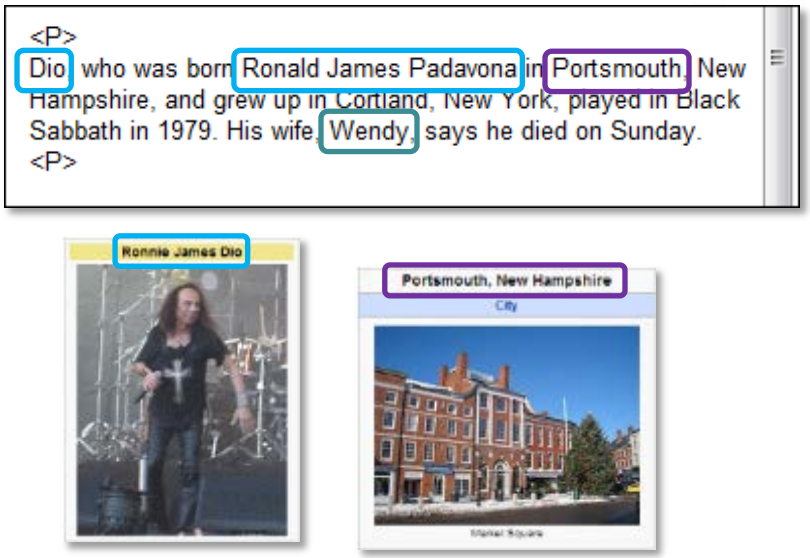
TED&L: Gold Standard Data Production

<P>
Dio, who was born Ronald James Padavona in Portsmouth, New Hampshire, and grew up in Cortland, New York, played in Black Sabbath in 1979. Ronnie's wife, Wendy, says he died on Sunday.
<P>

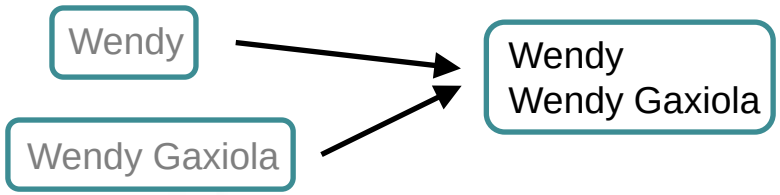
- Five entity types
  - PERs, ORGs, GPEs, FACs, LOCs
- Two mention types
  - Names and nominals
- Titles
  - Annotated to help distinguish PER nominal mentions
  - "the president [PER.NOM] signed a bill today“
  - "President[TTL] Clinton [PER.NAM] made a speech today"

TED&L: Gold Standard Data Production

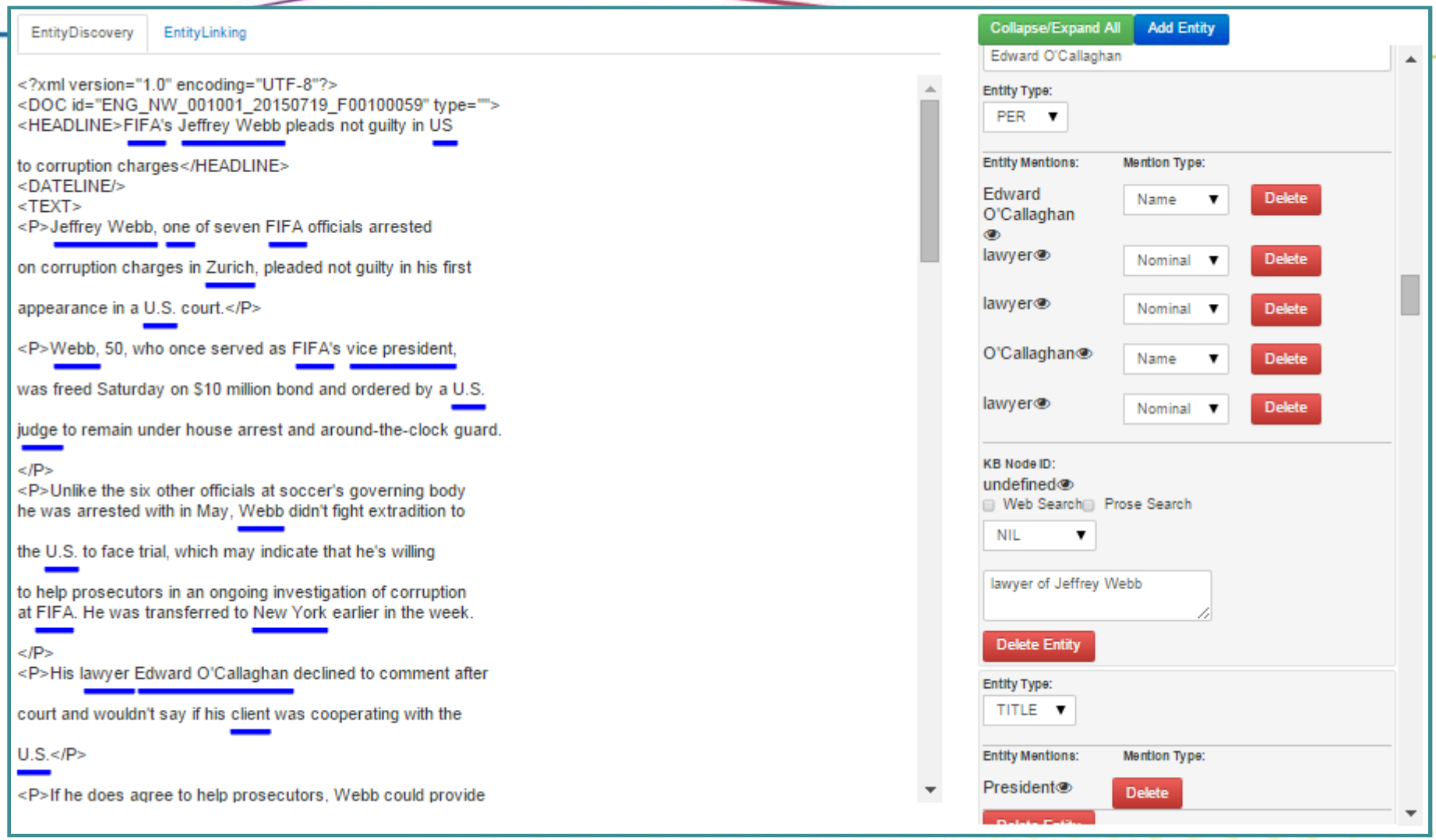
- KB Linking
  - Review ref document and search KB for matching node
  - Multiple entities viewed together for quicker linking



- NIL Coreference
  - NIL queries (no KB match) require manual co-reference annotation



LDC's TED&L/ED GUI

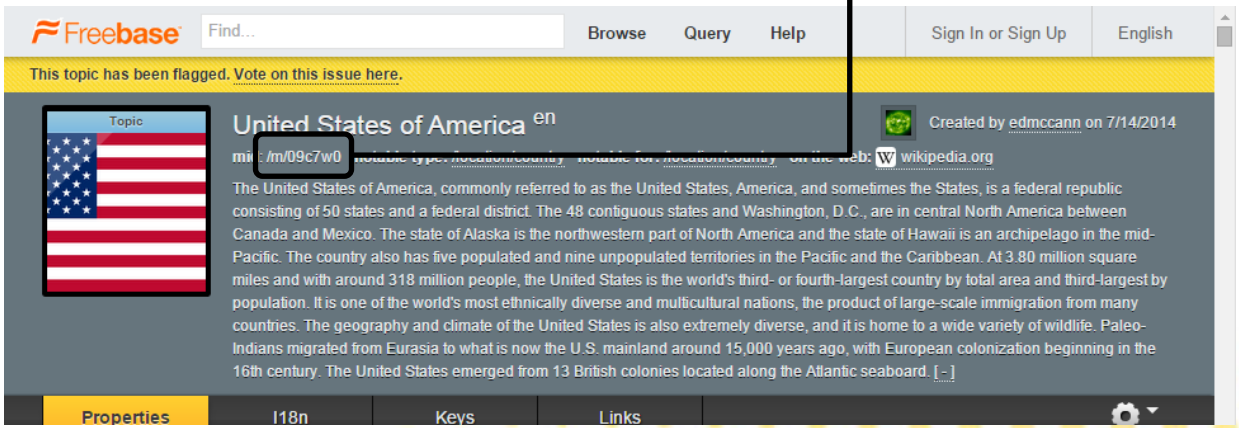


TED&L: Gold Standard Data Production

- Three types of annotators for within-doc kits
  - Monolingual English
  - Bilingual Chinese/English
  - Bilingual Spanish/English
- One English annotator for cross-doc coreference
  - Using English descriptions about the referents for each non-English cluster
    - Descriptions produced manually by original annotators during within-doc annotation

TED&L: KB-linking workaround

- Developed multiple indices for searching human-readable KB
  - None produced workable search results
    - Search for "united states" produced US page as 650th result
- Workaround
  - Annotators searched Freebase online and copied entity IDs
  - IDs not appearing in BaseKB converted to NILs during processing
- Currently working with NIST to create a better, more sustainable approach for 2016



TED&L: Data Volumes

	Training data (LDC2015E75)	Eval data (LDC2015E103)
Total mentions	**30,838**	**32,533**
ENG	13,545	15,645
CMN	13,116	11,066
SPA	4,177	5,822
Total equivalence classes	**5,744**	**7,235**
ENG	2,702	3,190
CMN	1,827	2,139
SPA	739	1,363
ENG/CMN	170	159
ENG/SPA	96	123
CMN/SPA	38	38
ENG/CMN/SPA	172	223

New 2015 Resources

Catalog ID	Corpus Title	Size
LDC2015E42	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Knowledge Base	1 knowledge base
LDC2015E43	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Knowledge Base Entries Creation Algorithm	1 algorithm
LDC2015E44	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Pilot Gold Standard Knowledge Base Links V1.1	686 mentions
LDC2015E61	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Pilot Source Corpus	15 documents
LDC2015E75	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Training Data V2.1	30838 mentions
LDC2015E93	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Evaluation Source Corpus V2.0	500 documents
LDC2015E102	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Evaluation Queries V1.2	32,533 queries
LDC2015E103	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Evaluation Gold Standard Entity Mentions & Knowledge Base Links	32,533 mentions
LDC2015E72	TAC KBP 2015 English Cold Start Entity Discovery Sample Data	162 mentions
LDC2015E76	TAC KBP 2015 English Cold Start Evaluation Queries V2.0	2539 queries
LDC2015E77	TAC KBP 2015 English Cold Start Evaluation Source Corpus V2.0	49124 documents
LDC2015E80	TAC KBP 2015 English Cold Start Evaluation Queries and Manual Run	2218 responses
LDC2015E81	TAC KBP 2015 English Cold Start Entity Discovery Evaluation Gold Standard Entity Mentions V1.2	8110 mentions
LDC2015E100	TAC KBP 2015 English Cold Start Evaluation Assessment Results V3.1	30,678 assessments