CLAMS: Computational Linguistic Applications for Multimedia Services

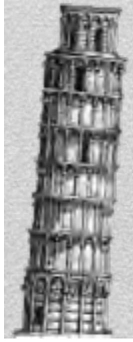
James Pustejovsky

Brandeis University

October 8, 2018

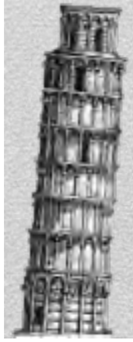
CLARIN Workshop

Pisa, Italy



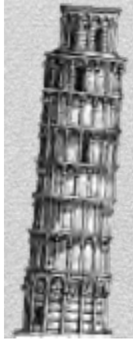
Collaboration

- *Brandeis*:
  - *James Pustejovsky*
  - *Kyeongmin Rim*
  - *Kelley Lynch*
  - *Ken Lai*
  - *Marc Verhagen*
- *WGBH Archives*:
  - *Karen Cariani*
  - *Casey Davis Kaufman*
  - *Sadie Roosa*



Talk Outline

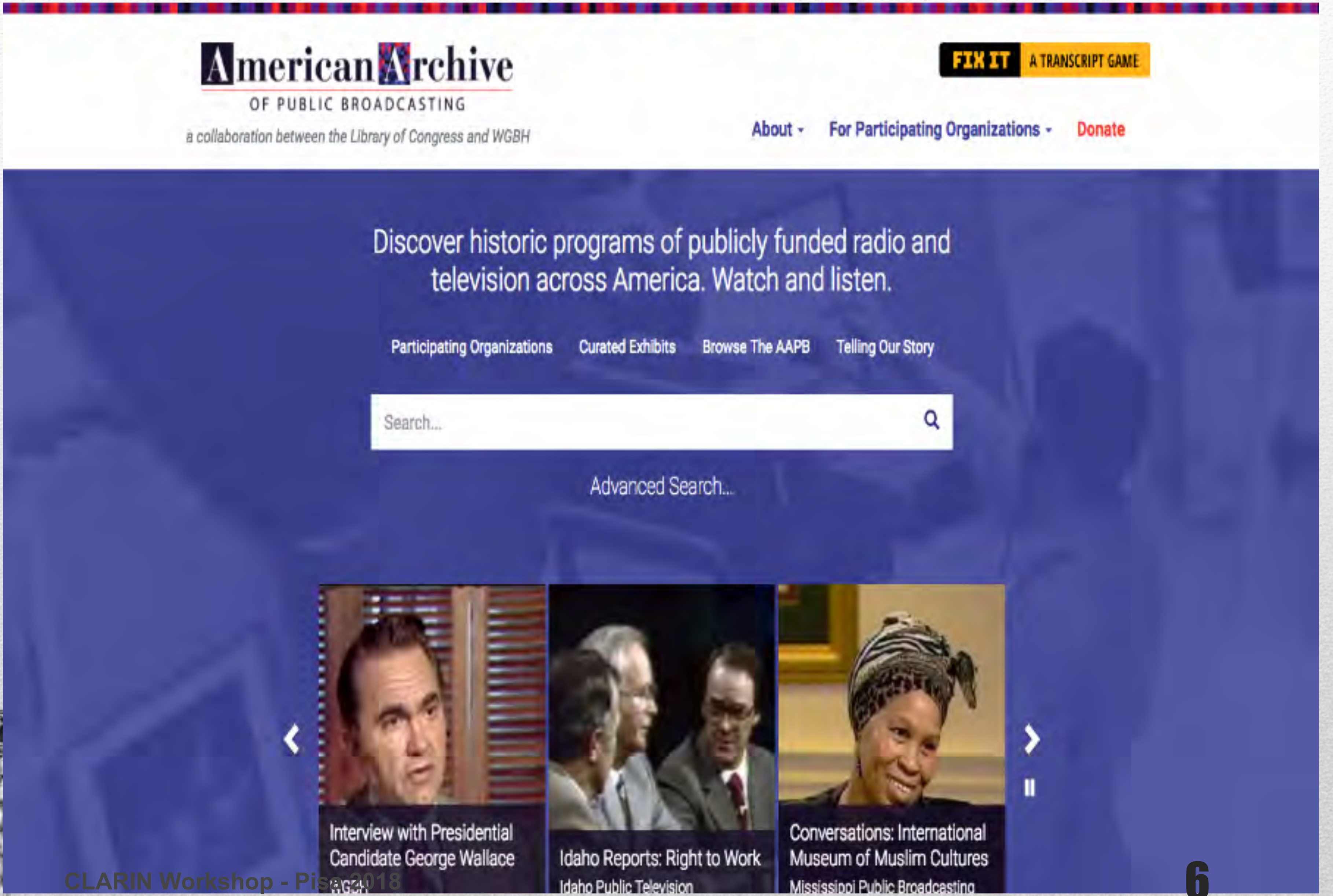
- Motivation - discoverability through enhanced metadata
  - American Archive and WGBH Corpus
- Background- LAPPS, HathiTrust, WebLicht, LIF, TCF
- CLAMS - Interoperability for video, images, and audio
- Architecture and Interchange Formats
- CLAMS Apps
  - Bars-tones Filtering
  - Text-in-image recognition
  - Forced alignment
  - Chaptering
  - Event Classification and Localization
- Scenario demo



AAPB

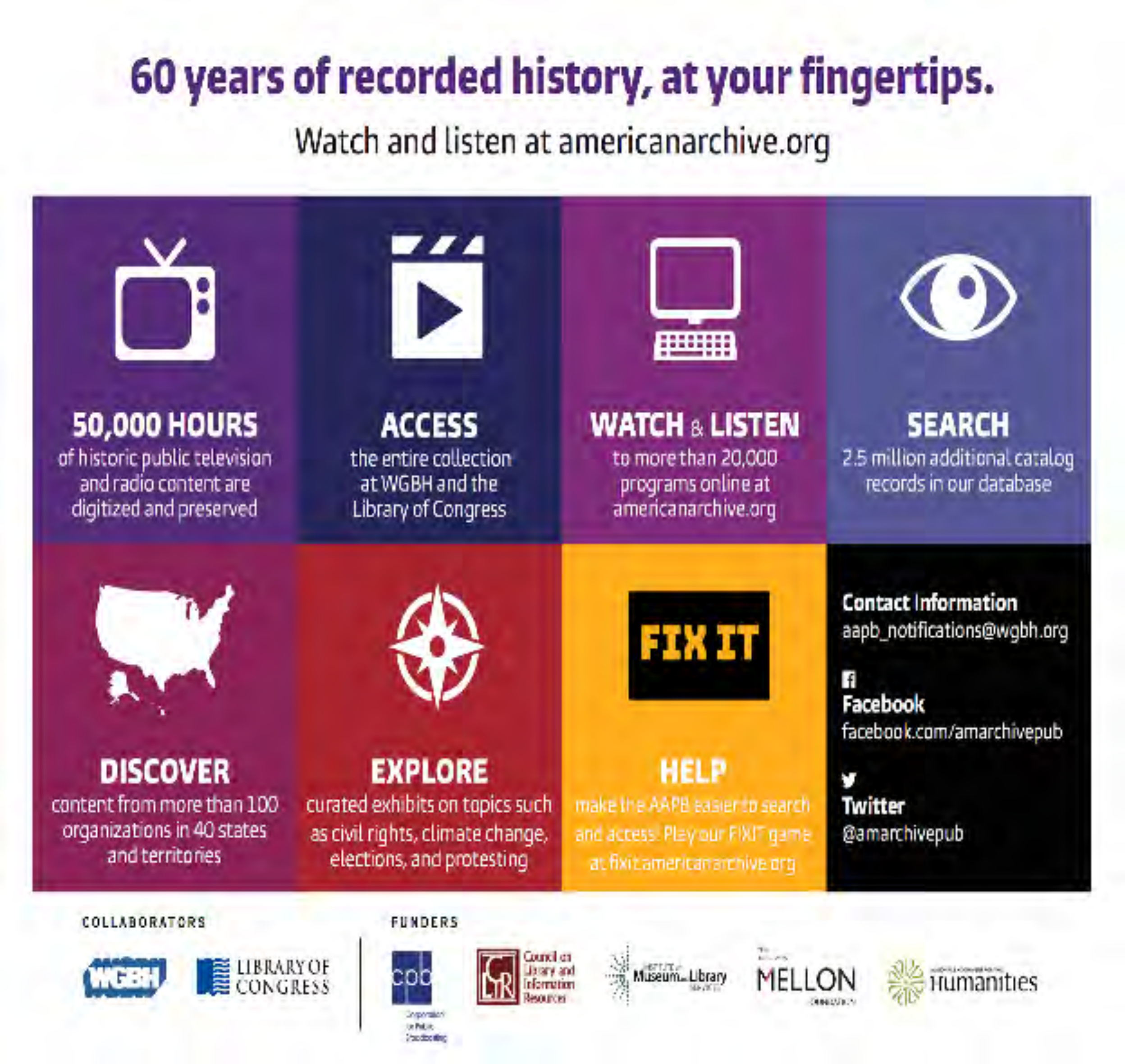
American Archive

OF PUBLIC BROADCASTING



American Archive
OF PUBLIC BROADCASTING
a collaboration between the Library of Congress and WGBH
FIX IT A TRANSCRIPT GAME
About
For Participating Organizations
Donate
Participating Organizations
Curated Exhibits
Browse The AAPB
Telling Our Story
Search...
Advanced Search...
Interview with Presidential Candidate George Wallace
WGBH
Idaho Reports: Right to Work
Idaho Public Television
Conversations: International Museum of Muslim Cultures
Mississippi Public Broadcasting

Discover historic programs of publicly funded radio and television across America. Watch and listen.



60 years of recorded history, at your fingertips.

Watch and listen at americanarchive.org

50,000 HOURS
of historic public television and radio content are digitized and preserved

ACCESS
the entire collection at WGBH and the Library of Congress

WATCH & LISTEN
to more than 20,000 programs online at americanarchive.org

SEARCH
2.5 million additional catalog records in our database

DISCOVER
content from more than 100 organizations in 40 states and territories

EXPLORE
curated exhibits on topics such as civil rights, climate change, elections, and protesting

FIX IT

HELP
make the AAPB easier to search and access. Play our FIXIT game at fixit.americanarchive.org

Contact Information
aapb_notifications@wgbh.org

Facebook
facebook.com/amarchivepub

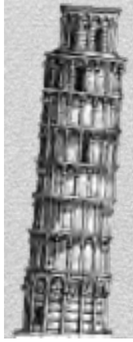
Twitter
@amarchivepub

COLLABORATORS

WGBH LIBRARY OF CONGRESS

FUNDERS

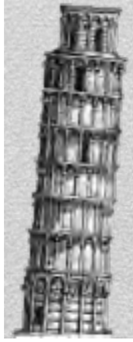
cpb Corporation for Public Broadcasting · Council on Library and Information Resources · Museum and Library · MELLON · Humanities



LAPPS GRID

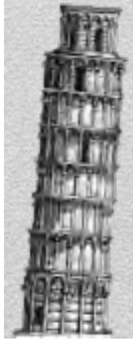
The LAPPS Grid Project

- Collaborative effort among US partners
  - Brandeis University - James Pustejovsky
  - Vassar College – Nancy Ide
  - Carnegie-Mellon University – Eric Nyberg
  - Linguistic Data Consortium (U. of Pennsylvania) – Chris Cieri
- Funded by the US National Science Foundation
- Builds on
  - foundation laid in several projects
    - SILT, The Language Grid, PANACEA, LinguaGrid… momentum toward a comprehensive network of web services and resources within the NLP community



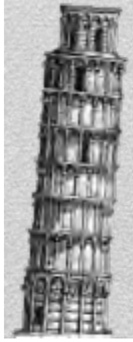
The LAPPS Grid

- A framework to
  - enable language service discovery, composition, and reuse
    - For both NLP researchers and others (who may use pre-developed composite services)
  - promote sustainability, manageability, usability, and interoperability of NLP components
- Based on the service-oriented architecture (SOA)
  - Web-oriented version of the “pipeline” architecture for sequencing loosely-coupled linguistic analyses



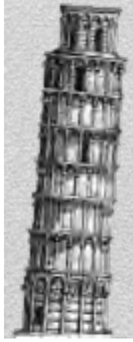
Overall Goals

- Design, develop, and promote a Language Application Grid based on Service Grid Software
  - Support **development and deployment of integrated natural language applications**
  - Enable federation of grids and services
- Provide an open advancement (OA) framework for component- and application-based evaluation
- Provide access to language resources for members of the NLP community as well as researchers in a wide range of social science and humanities disciplines
- Enable easy navigation through licensing issues
- Actively promote adoption, use, and community involvement with the LAPPS Grid
- Actively pursue creation of an interoperable global network of grids and frameworks



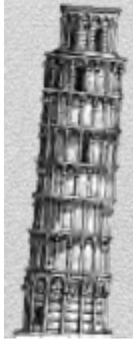
Functionality

- Provides access to
  - basic NLP processing tools
  - language resources such as mono- and multi-lingual corpora and lexicons
- Enables pipelining tools to create custom NLP applications and "black box" composite services
- Ultimately a community-based project
  - Services contributed by members of the community
  - Existing service repositories and grids federated to enable universal access



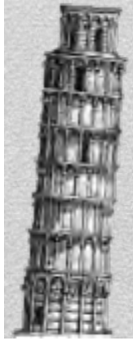
Transformative aspects

- Orchestrates access to and deployment of language resources and processing functions available from servers around the globe
- Enables users to add their own language resources, services, and even service grids
- Provides a critical missing layer of functionality for NLP
  - Current frameworks (e.g., GATE, UIMA) do not provide general support for service discovery, composition, and reuse
    - Communication among tools based on a specific internal format (e.g. UIMA CAS)
    - LAPPS Grid enables calling tools and pipelines within GATE, UIMA, etc. as services themselves
      - Thus interoperable with all other LAPPS Grid services

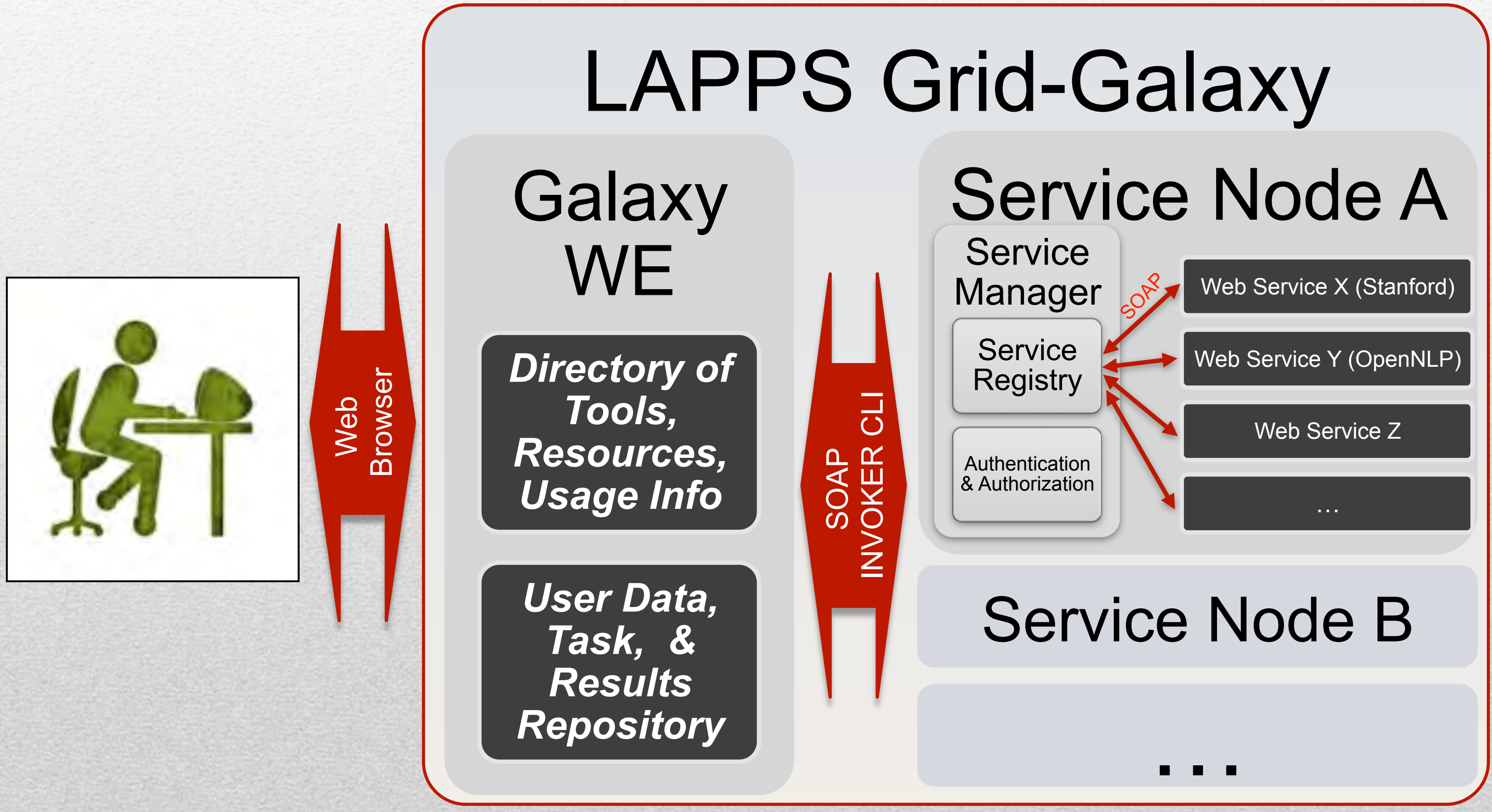


Overall Architecture

- Based on the Open Service Grid Initiative's Service Grid Server Software
  - Developed by the National Institute of Information and Communications Technology (NICT) in Japan
  - Used to implement Kyoto University's Language Grid
    - Also used for several Asian grids, soon-to-come ELRA Grid



LAPPS-Galaxy Architecture



LAPPS Grid-Galaxy

Galaxy WE

Directory of Tools, Resources, Usage Info

User Data, Task, & Results Repository

Service Node A

Service Manager

Service Registry

Authentication & Authorization

Web Service X (Stanford)

Web Service Y (OpenNLP)

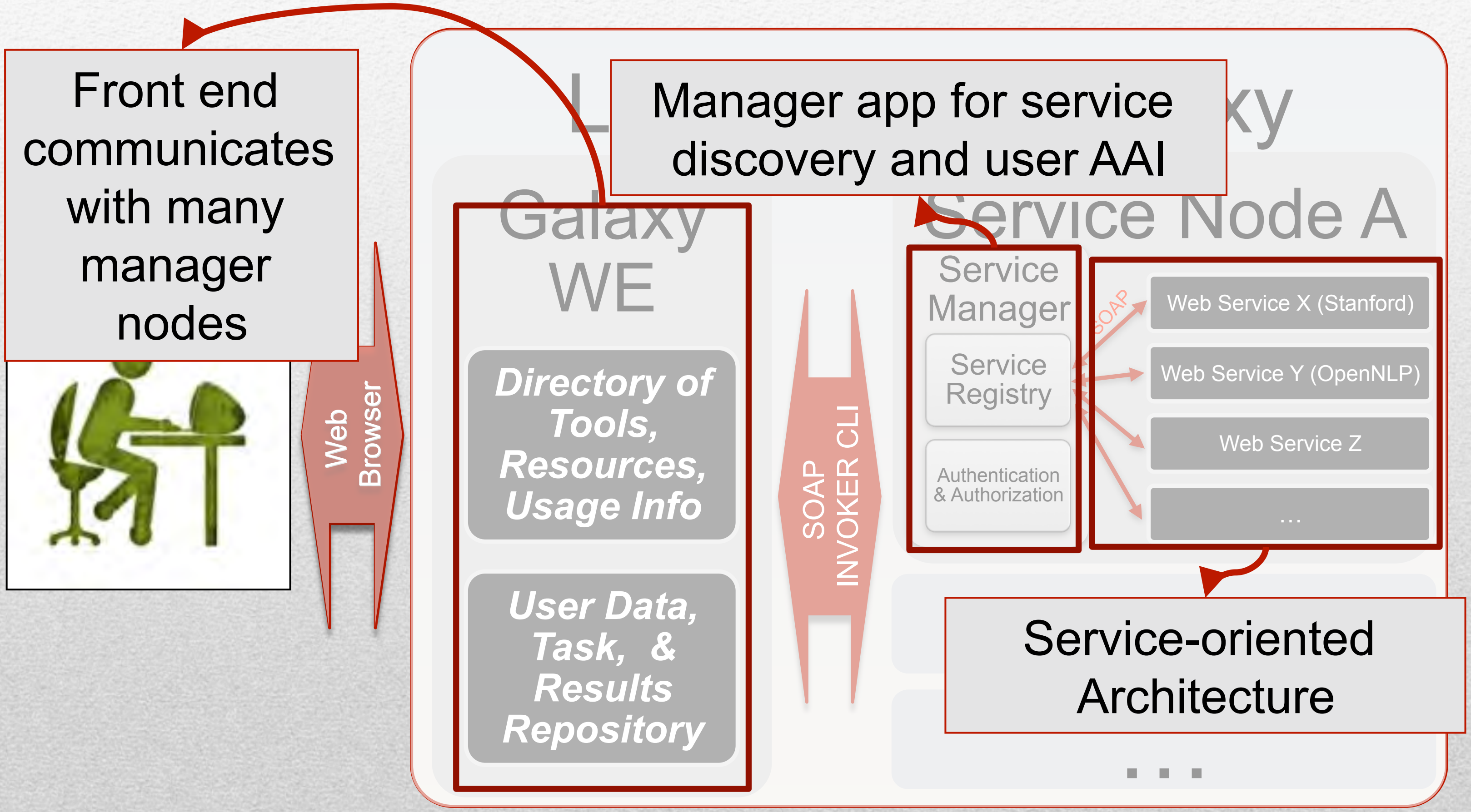
Web Service Z

…

Service Node B

. . .

LAPPS-Galaxy Architecture



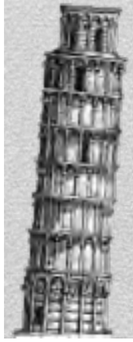
Front end communicates with many manager nodes

Manager app for service discovery and user AAI

Service-oriented Architecture

Extension of Service Grid Software

- Enhances capabilities for composition of tool and resource chains
- Provides sophisticated evaluation services
- Implements a dynamic licensing system for handling license agreements on the fly
- Provides the option to run services locally or in the cloud, with high-security technology to protect sensitive information
- Improves data delivery services
- Enables access to grids other than those based on the Service Grid technology
- Provides user-friendly, transparent facilities for wrapping user-provided services



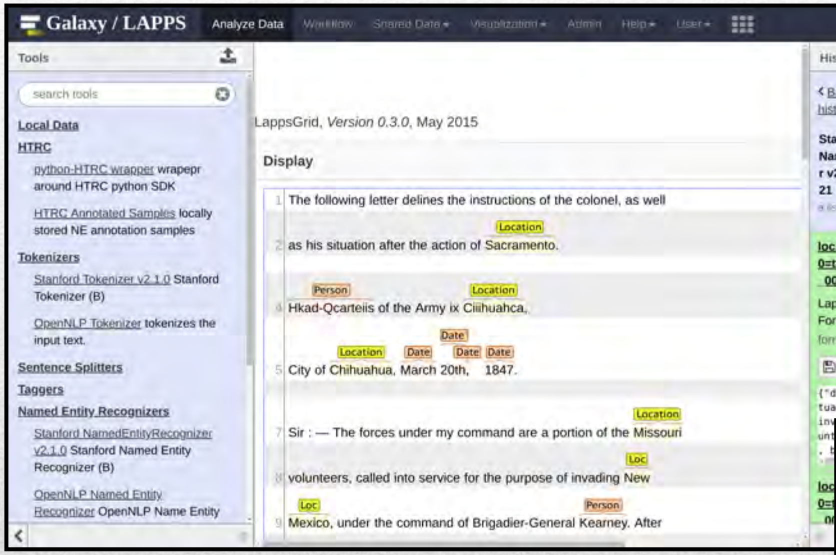
Galaxy as Front End

- The LAPPS Grid adopted the GALAXY workflow engine as a front end for construction of pipelines etc.

Galaxy

http://galaxyproject.org

Galaxy Workflow Engine



Local Data
HTRC
python-HTRC wrapper wrapepr around HTRC python SDK
HTRC Annotated Samples locally stored NE annotation samples
Tokenizers
Stanford Tokenizer v2.1.0 Stanford Tokenizer (B)
OpenNLP Tokenizer tokenizes the input text.
Sentence Splitters
Taggers
Named Entity Recognizers
Stanford NamedEntityRecognizer v2.1.0 Stanford Named Entity Recognizer (B)
OpenNLP Named Entity Recognizer OpenNLP Name Entity

LappsGrid, *Version 0.3.0*, May 2015

Display

1 The following letter delines the instructions of the colonel, as well
2 as his situation after the action of Sacramento. (Location)
4 Hkad-Qcarteiis of the Army ix Ciiihuahca. (Person, Location)
5 City of Chihuahua, March 20th, 1847. (Location, Date, Date, Date, Date)
7 Sir : — The forces under my command are a portion of the Missouri (Location)
8 volunteers, called into service for the purpose of invading New (Loc)
9 Mexico, under the command of Brigadier-General Kearney. After (Loc, Person)

Back to Unnamed history

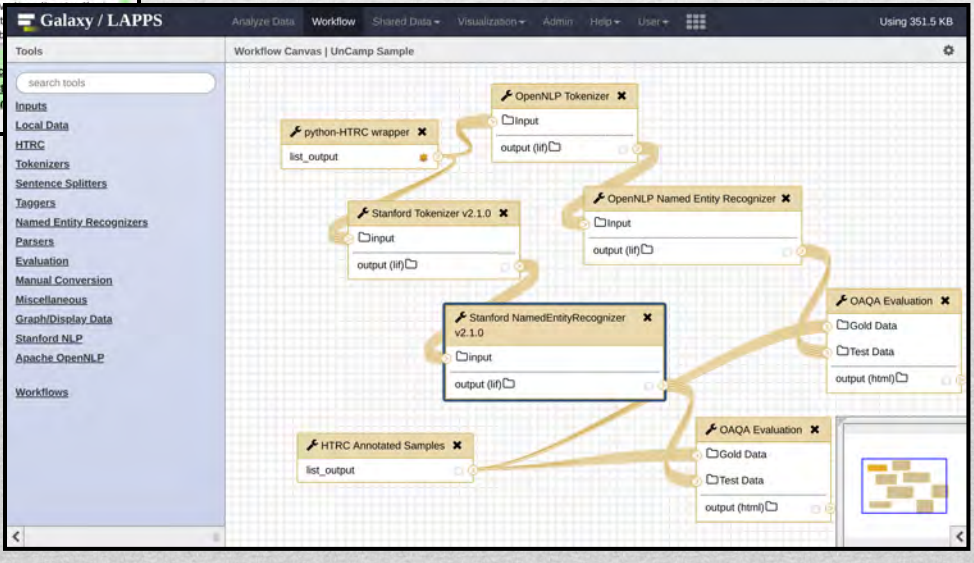
Stanford NamedEntityRecognizer v2.1.0 on collection 21
a list of datasets

loc.ark+=13960=t8ff3w38m_00000177
Lapps Interchange Format (LIF)
format: lif, database: ?

```
{"discriminator":"http
tuation after the acti
```

Galaxy Web Interface
(http://galaxy.lappsgrid.org/)

Galaxy Workflow Editor



Interoperability

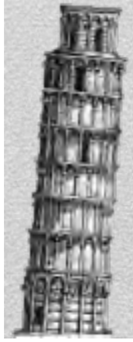
- Basic web service interoperability handled by SOAP/WSDL
- LAPPS Interchange Format (LIF)
  - format that allows services to exchange more detailed information
  - Syntactic interoperability
    - handled by JSON-LD
    - enforced by the LIF JSON schema
  - Semantic interoperability
    - enhanced by using the Linked Data aspect of JSON-LD to link to the LAPPS Web Services Exchange Vocabulary

Why JSON-LD

- Lightweight, text-based, language-independent data interchange format
- Based on the W3C Resource Definition Framework (RDF)
- Trivially mappable to and from other graph-based formats such as ISO LAF/GrAF , UIMA CAS
- Enables services to reference categories and definitions in web-based repositories and ontologies or any concept defined at a given URI

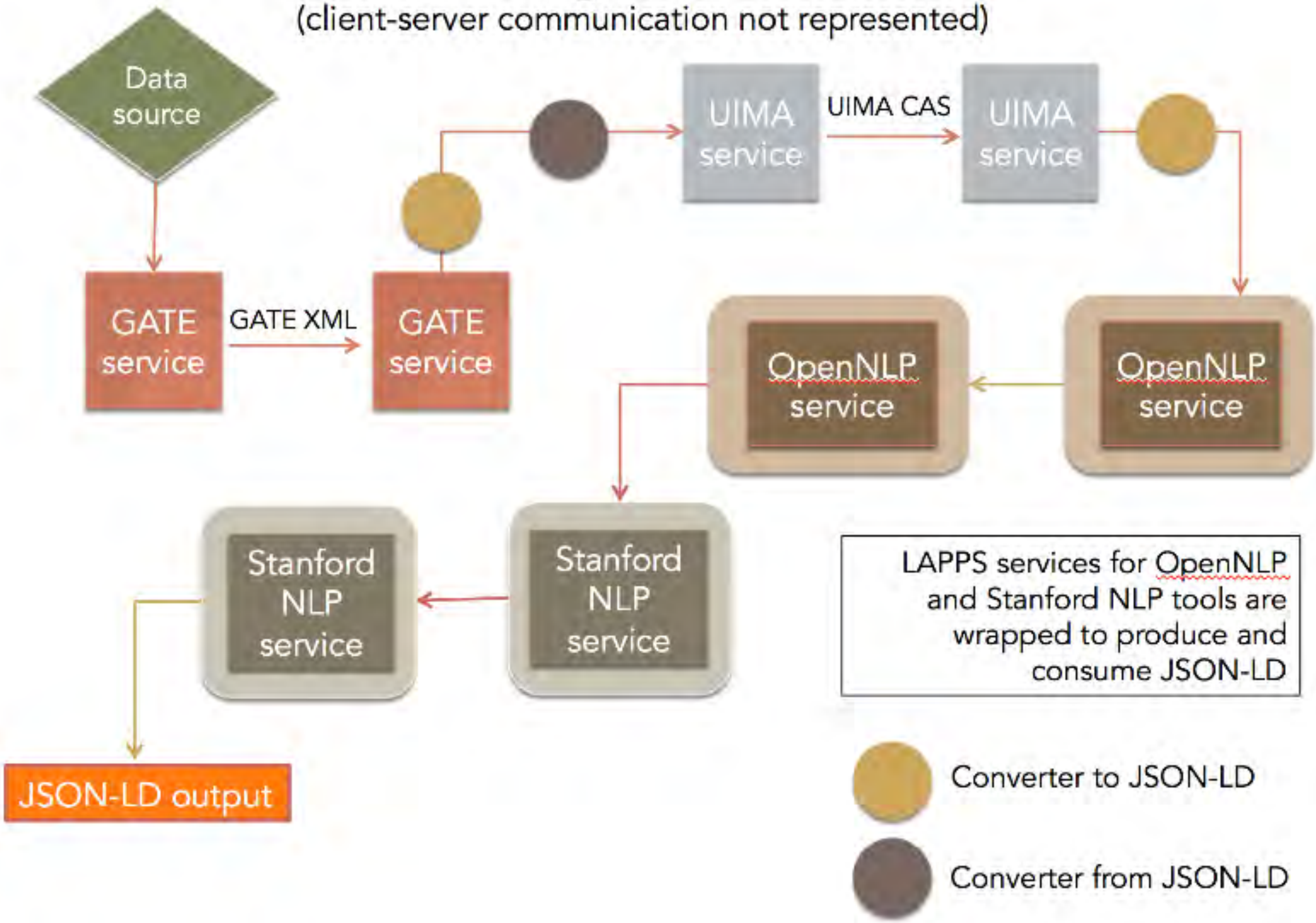
Web Service Communication in LAPPS

- Each service in the LAPPS Grid publishes metadata:
  - a discriminator (type) : tells how to interpret the payload
  - a payload (typically a utf-8 string)
- LAPPS uses JSON-LD as its standard format for the payload
  - Converters to and from JSON-LD for services that deliver in other formats
  - Some LAPPS services are wrapped to produce and consume JSON-LD



Logical flow

(client-server communication not represented)



LAPPS services for OpenNLP and Stanford NLP tools are wrapped to produce and consume JSON-LD

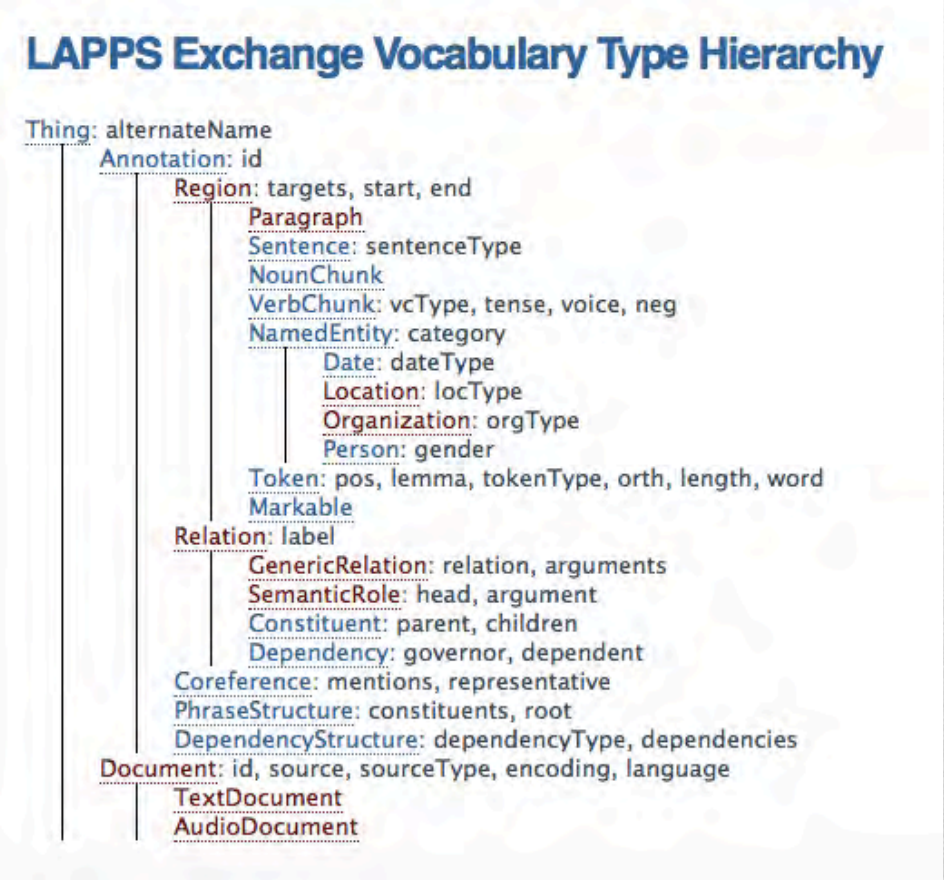
● Converter to JSON-LD

● Converter from JSON-LD

LAPPS Web Service Exchange Vocabulary

- No accepted standard for module description or input/output interchange in the language application domain currently exists
- LAPPS Web Service Exchange Vocabulary (WS-EV)
  - Specifies a terminology for a core of linguistic objects and features exchanged among NLP tools that consume and produce linguistically annotated data
  - May help address a need within the community to identify a standard terminology and indicate the relations among them
  - Linked wherever possible to existing repositories such as ISOCat (CLARIN Concept Repository), schema.org, FoLiA categories, etc.

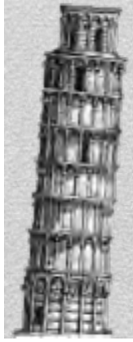
LAPPS Web Service Exchange Vocabulary



LAPPS Exchange Vocabulary Type Hierarchy

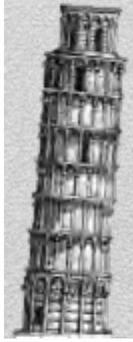
- Thing: alternateName
  - Annotation: id
    - Region: targets, start, end
      - Paragraph
      - Sentence: sentenceType
      - NounChunk
      - VerbChunk: vcType, tense, voice, neg
      - NamedEntity: category
        - Date: dateType
        - Location: locType
        - Organization: orgType
        - Person: gender
      - Token: pos, lemma, tokenType, orth, length, word
      - Markable
    - Relation: label
      - GenericRelation: relation, arguments
      - SemanticRole: head, argument
      - Constituent: parent, children
      - Dependency: governor, dependent
    - Coreference: mentions, representative
    - PhraseStructure: constituents, root
    - DependencyStructure: dependencyType, dependencies
  - Document: id, source, sourceType, encoding, language
    - TextDocument
    - AudioDocument

http://vocab.lappsgrid.org/



JSON-LD and WS-EV

- References in JSON-LD representation point to URIs providing **definitions** for specific linguistic categories in the WS-EV
- Also point to **documentation** for processing software and rules for processes such as tokenization, entity recognition, etc.
  - Often left unspecified in annotated resources
  - Not required for web service exchange in the LAPPS Grid
  - **BUT** inclusion of such references can contribute to better replication and evaluation of results in the field
  - **Promote best practice**!

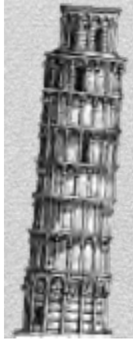


LAPPS IN THE HATHITRUST

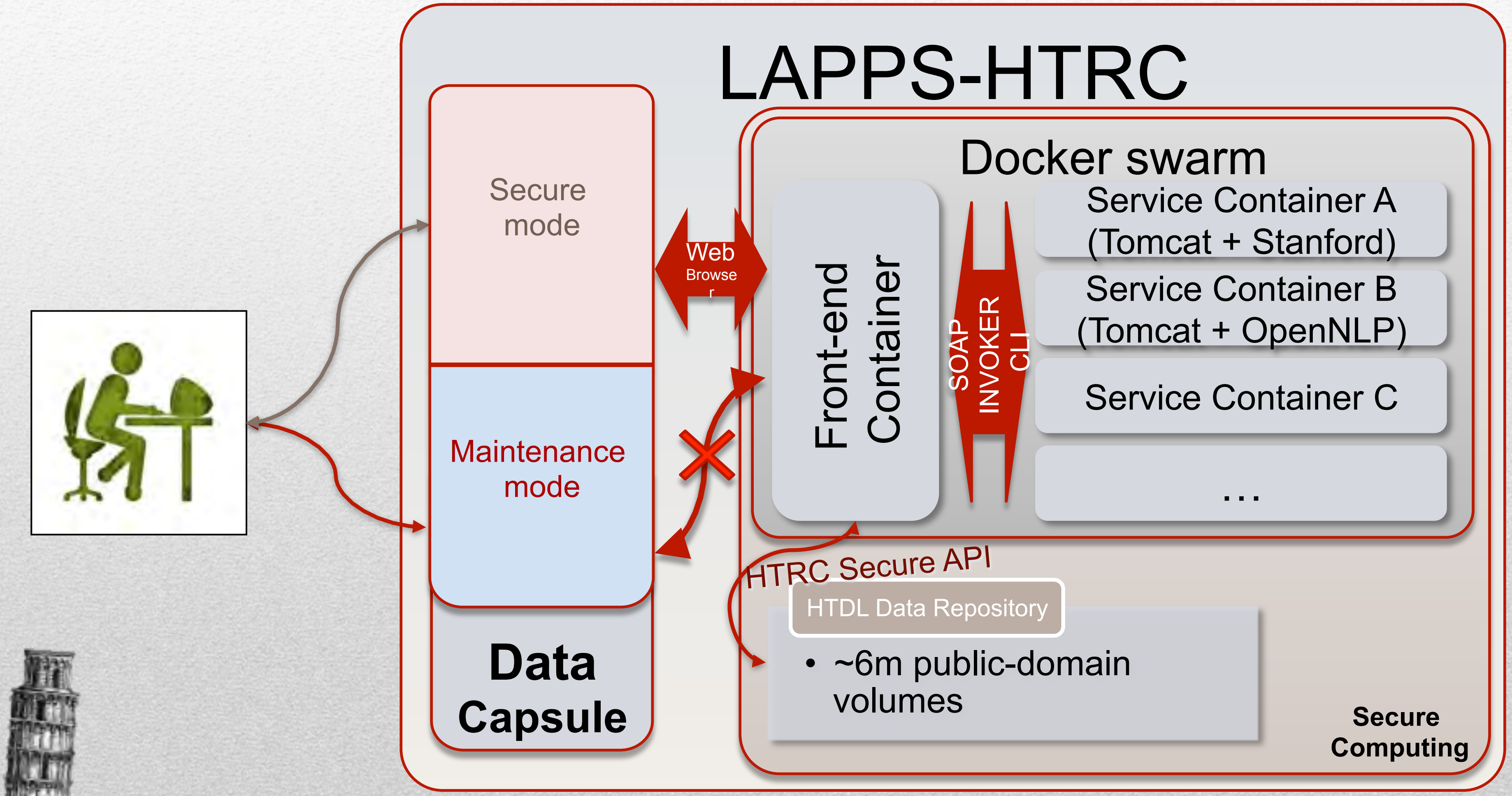
University of Illinois
Indiana University
Brandeis University

LAPPS-HTRC Collaboration

- HathiTrust Digital Library
  - The HathiTrust is a consortium of members that steward the over 15 million volumes of digitized content from research libraries across the world.
  - Long-term preservation and access services for public domain and in copyright content (from a variety of sources, including Google, the Internet Archive, Microsoft and more)
- HathiTrust Research Center (HTRC)
  - Providing means for researchers to analyze large swaths of the 15+ million volumes of HathiTrust drawing on computational resources and tools



LAPPS-HTRC Collaboration



LAPPS-HTRC Collaboration

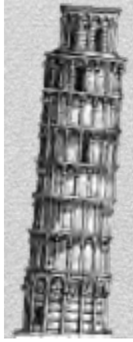
- AAI by HTRC secure computing environment (Data Capsule)
- Removed the manager app: the Galaxy WE front end directly invokes language services
- Front end and services are wrapped in individual docker containers and orchestrated as a swarm of virtual servers
- LAPPS-Galaxy is deployed as a docker swarm on the physical machine that can access in-copyright data repository

CL Tools for HTRC Data Capsules

- Enhance search and discovery across the library by complementing traditional volume-level bibliographic metadata with new metadata
- Creation of Linked Open Data resources to help scholars find, select, integrate and disseminate a wider range of data as part of their scholarly analysis life-cycle
- Creation of a set of pre-built DCs that incorporate tools commonly used by both the Digital Humanities and the CL communities that scholars can then customize to address their specific needs

LAPPS-HTRC in the Data Capsule

- The LAPPS-HTRC server runs on a separate server from DC hosts, outside the individual Capsules that are used for analysis. It operates under a similar set of firewall rules as HTRC DataAPI, working within the threat model of the Data Capsule service. In doing this, users have access to LAPPS in the secure mode of Data Capsule only.
- The LAPPS-HTRC has been deployed in the HTRC Services Development Platform, with access to the same entire HathiTrust public domain dataset as is served on the HTRC Production Platform, which is about 6 million volumes.
- Users can access the LAPPS-HTRC upon request. Users will be granted an account in the development portion of the HTRC Analytics website.

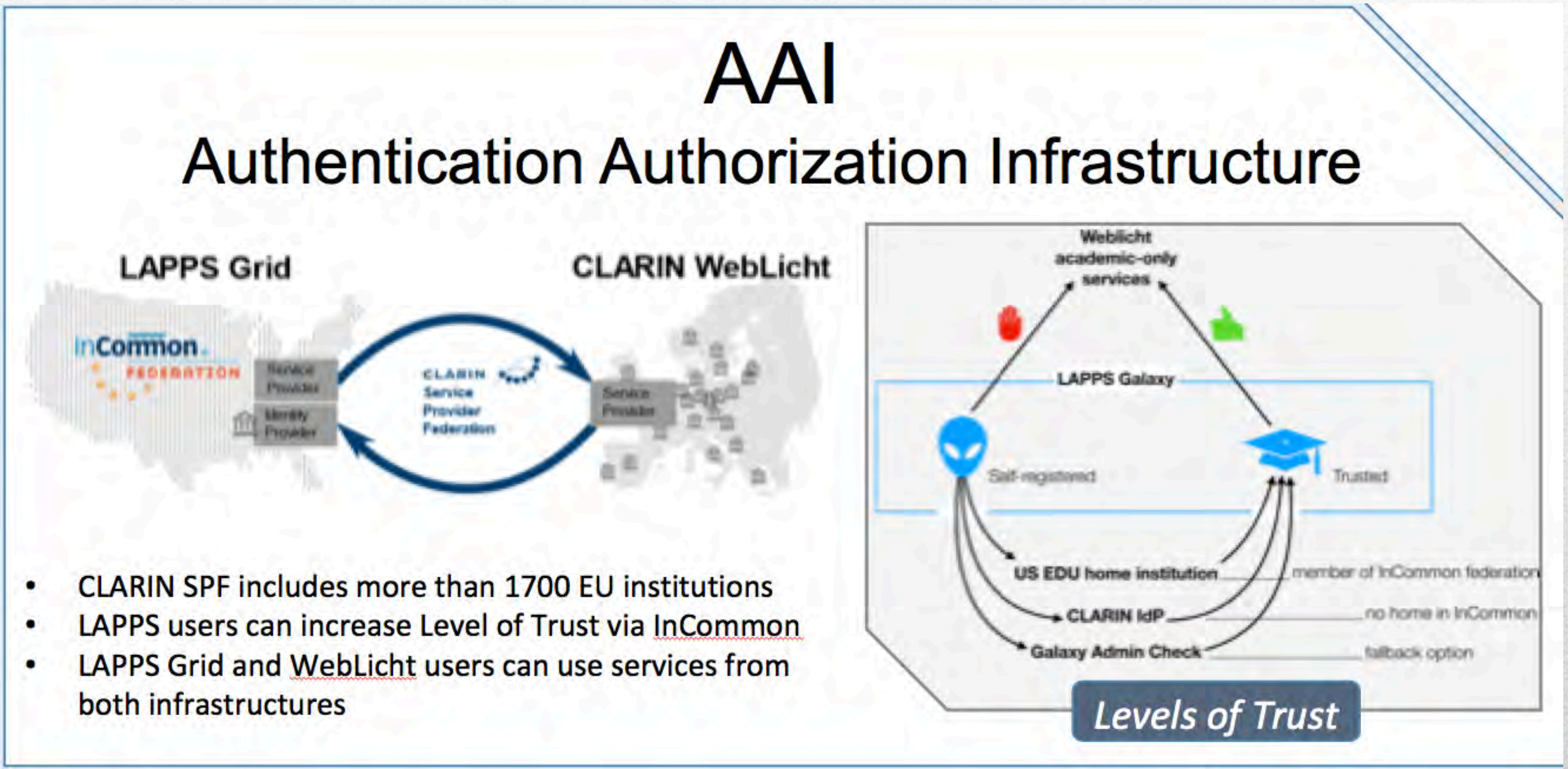


THE ANDREW W. MELLON FOUNDATION

BRIDGING CLARIN AND LAPPS

CLARIN-D - Tübingen
LINDAT/CLARIN – Prague
LAPPS – Brandeis and Vassar

Ensuring Authentication

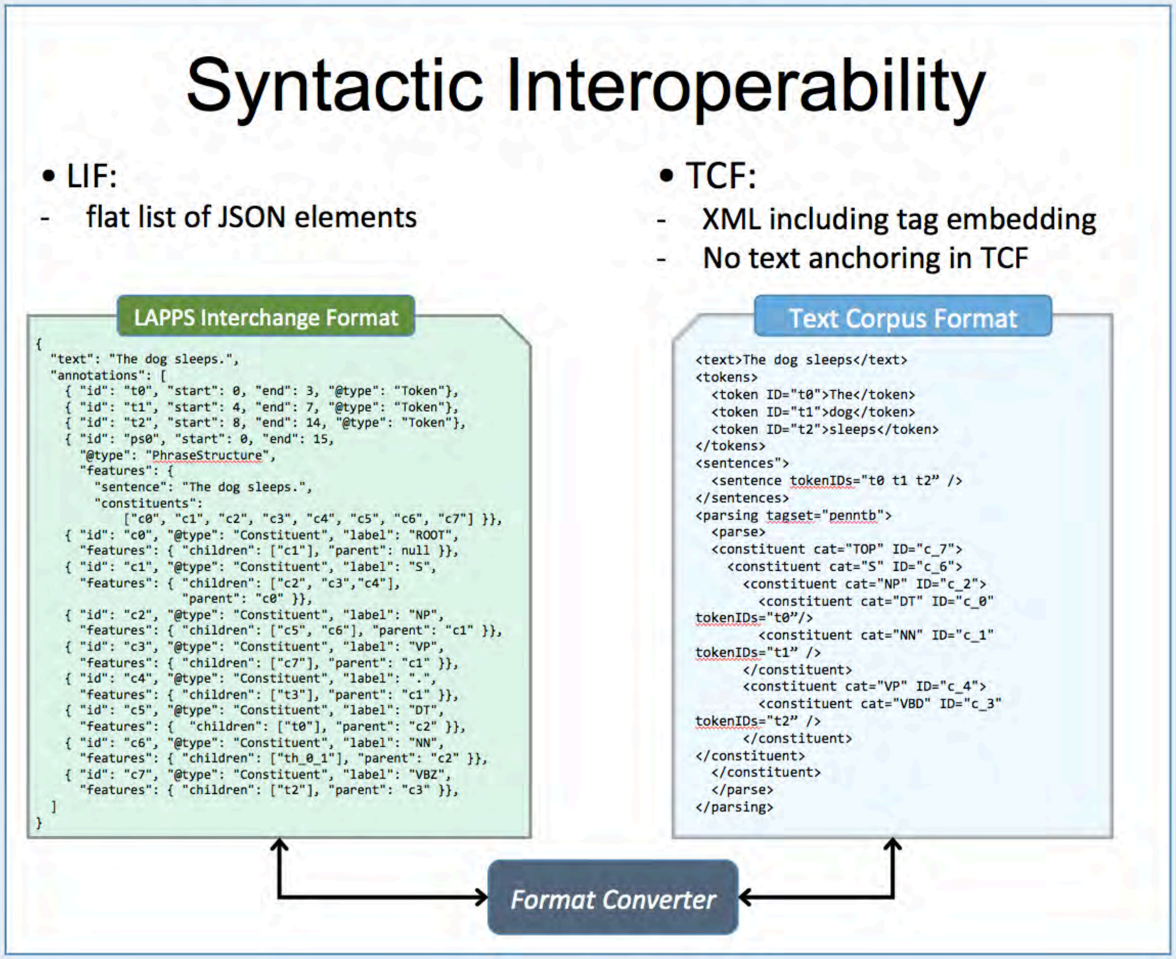


AAI

Authentication Authorization Infrastructure

- CLARIN SPF includes more than 1700 EU institutions
- LAPPS users can increase Level of Trust via InCommon
- LAPPS Grid and WebLicht users can use services from both infrastructures

Syntactic Interoperability



Syntactic Interoperability

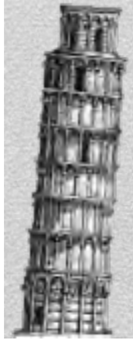
- LIF:
  - flat list of JSON elements
- TCF:
  - XML including tag embedding
  - No text anchoring in TCF

LAPPS Interchange Format

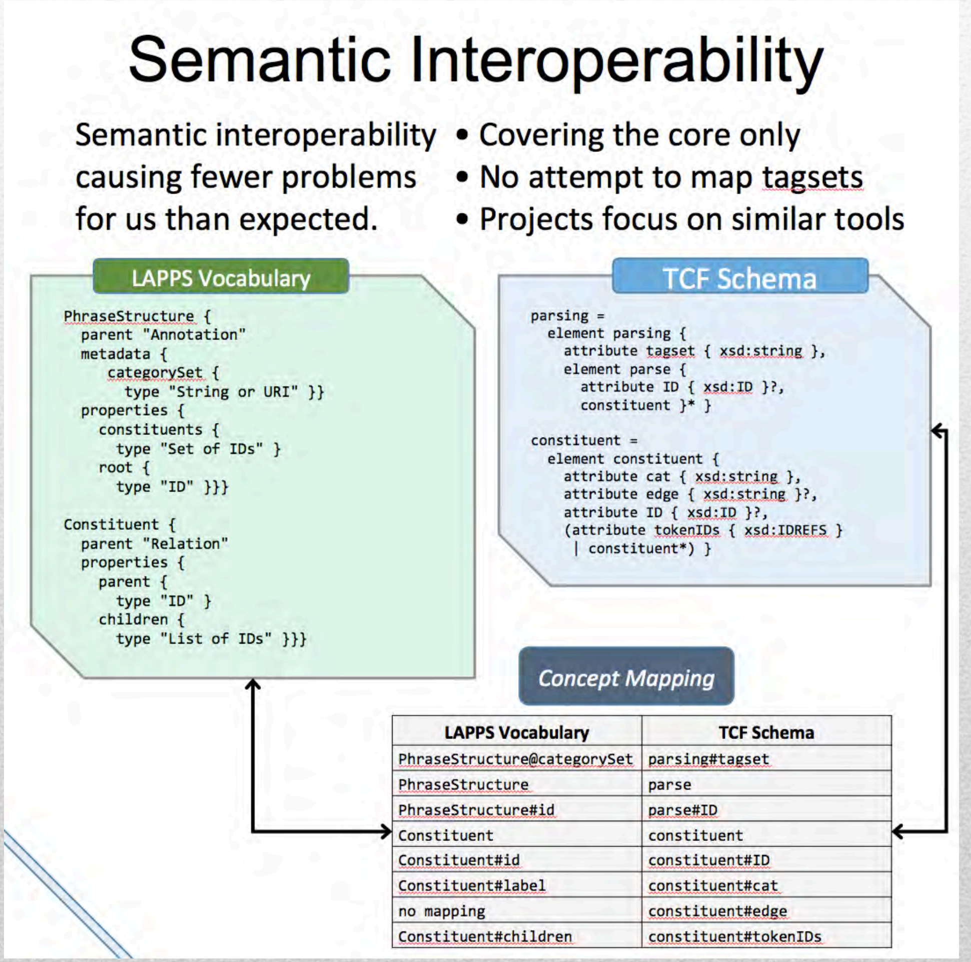
```
{
  "text": "The dog sleeps.",
  "annotations": [
    { "id": "t0", "start": 0, "end": 3, "@type": "Token"},
    { "id": "t1", "start": 4, "end": 7, "@type": "Token"},
    { "id": "t2", "start": 8, "end": 14, "@type": "Token"},
    { "id": "ps0", "start": 0, "end": 15,
      "@type": "PhraseStructure",
      "features": {
        "sentence": "The dog sleeps.",
        "constituents":
            ["c0", "c1", "c2", "c3", "c4", "c5", "c6", "c7"] }},
    { "id": "c0", "@type": "Constituent", "label": "ROOT",
      "features": { "children": ["c1"], "parent": null }},
    { "id": "c1", "@type": "Constituent", "label": "S",
      "features": { "children": ["c2", "c3","c4"],
                    "parent": "c0" }},
    { "id": "c2", "@type": "Constituent", "label": "NP",
      "features": { "children": ["c5", "c6"], "parent": "c1" }},
    { "id": "c3", "@type": "Constituent", "label": "VP",
      "features": { "children": ["c7"], "parent": "c1" }},
    { "id": "c4", "@type": "Constituent", "label": ".",
      "features": { "children": ["t3"], "parent": "c1" }},
    { "id": "c5", "@type": "Constituent", "label": "DT",
      "features": {  "children": ["t0"], "parent": "c2" }},
    { "id": "c6", "@type": "Constituent", "label": "NN",
      "features": { "children": ["th_0_1"], "parent": "c2" }},
    { "id": "c7", "@type": "Constituent", "label": "VBZ",
      "features": { "children": ["t2"], "parent": "c3" }},
  ]
}
```

Text Corpus Format

```
<text>The dog sleeps</text>
<tokens>
  <token ID="t0">The</token>
  <token ID="t1">dog</token>
  <token ID="t2">sleeps</token>
</tokens>
<sentences">
  <sentence tokenIDs="t0 t1 t2” />
</sentences>
<parsing tagset="penntb">
  <parse>
  <constituent cat="TOP" ID="c_7">
    <constituent cat="S" ID="c_6">
      <constituent cat="NP" ID="c_2">
        <constituent cat="DT" ID="c_0"
tokenIDs="t0”/>
        <constituent cat="NN" ID="c_1"
tokenIDs="t1” />
      </constituent>
      <constituent cat="VP" ID="c_4">
        <constituent cat="VBD" ID="c_3"
tokenIDs="t2” />
      </constituent>
</constituent>
  </constituent>
  </parse>
</parsing>
```



Semantic Interoperability



Semantic Interoperability

Semantic interoperability causing fewer problems for us than expected.

- Covering the core only
- No attempt to map tagsets
- Projects focus on similar tools

LAPPS Vocabulary

```
PhraseStructure {
  parent "Annotation"
  metadata {
     categorySet {
       type "String or URI" }}
  properties {
    constituents {
      type "Set of IDs" }
    root {
      type "ID" }}}

Constituent {
  parent "Relation"
  properties {
    parent {
      type "ID" }
    children {
      type "List of IDs" }}}
```

TCF Schema

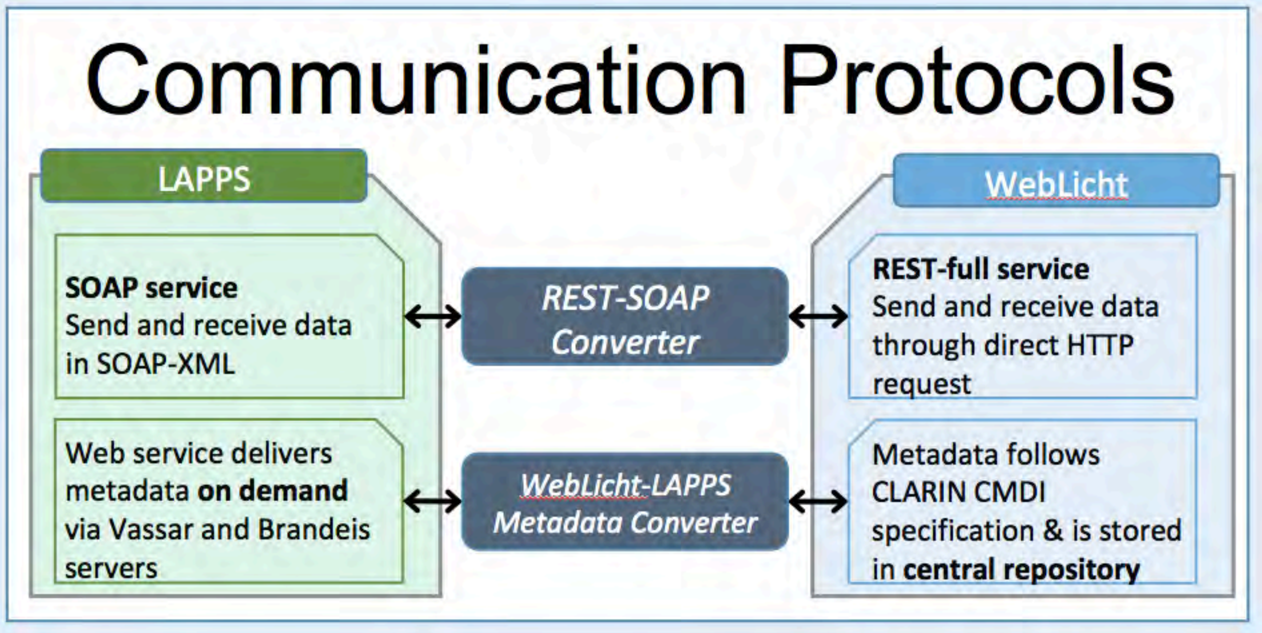
```
parsing =
  element parsing {
    attribute tagset { xsd:string },
    element parse {
      attribute ID { xsd:ID }?,
      constituent }* }

constituent =
  element constituent {
    attribute cat { xsd:string },
    attribute edge { xsd:string }?,
    attribute ID { xsd:ID }?,
    (attribute tokenIDs { xsd:IDREFS }
     | constituent*) }
```

Concept Mapping

LAPPS Vocabulary	TCF Schema
PhraseStructure@categorySet	parsing#tagset
PhraseStructure	parse
PhraseStructure#id	parse#ID
Constituent	constituent
Constituent#id	constituent#ID
Constituent#label	constituent#cat
no mapping	constituent#edge
Constituent#children	constituent#tokenIDs

Connecting different protocols



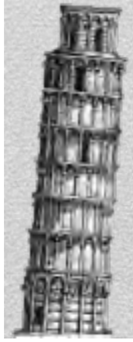
Communication Protocols

SOAP service Send and receive data in SOAP-XML

Web service delivers metadata **on demand** via Vassar and Brandeis servers

REST-full service Send and receive data through direct HTTP request

Metadata follows CLARIN CMDI specification & is stored in **central repository**



Integration: WebLicht and LAPPS

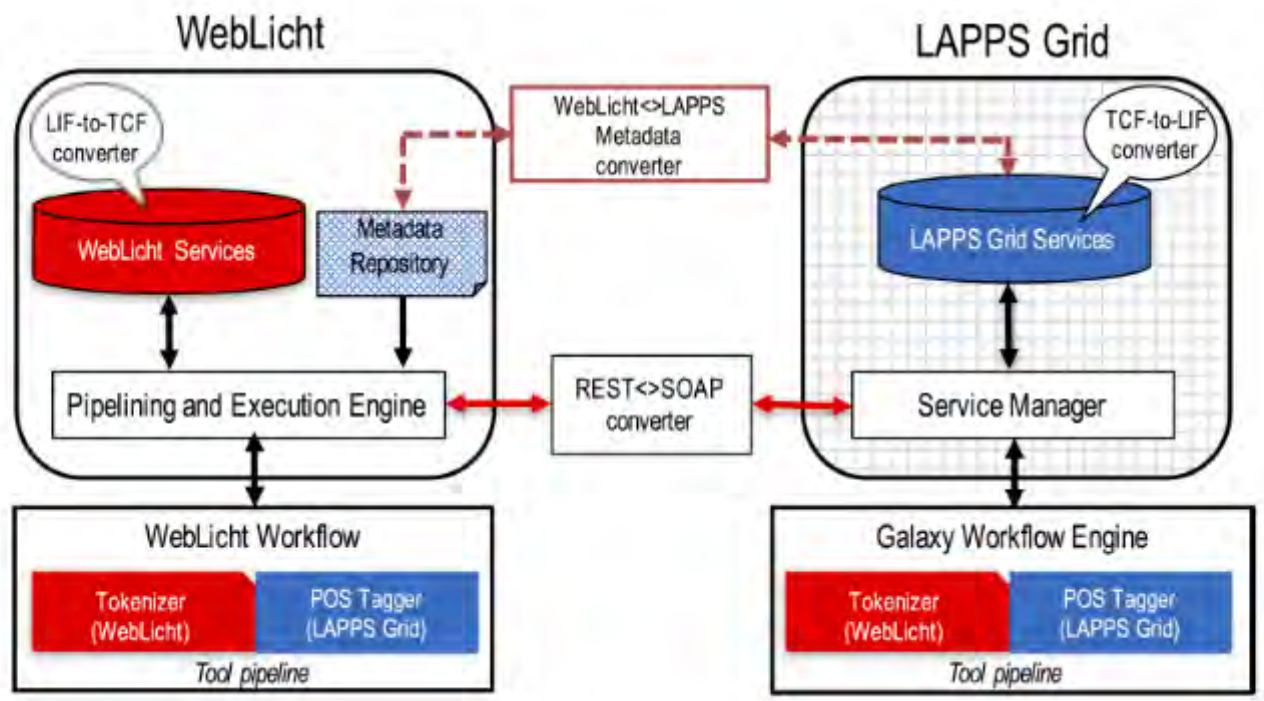
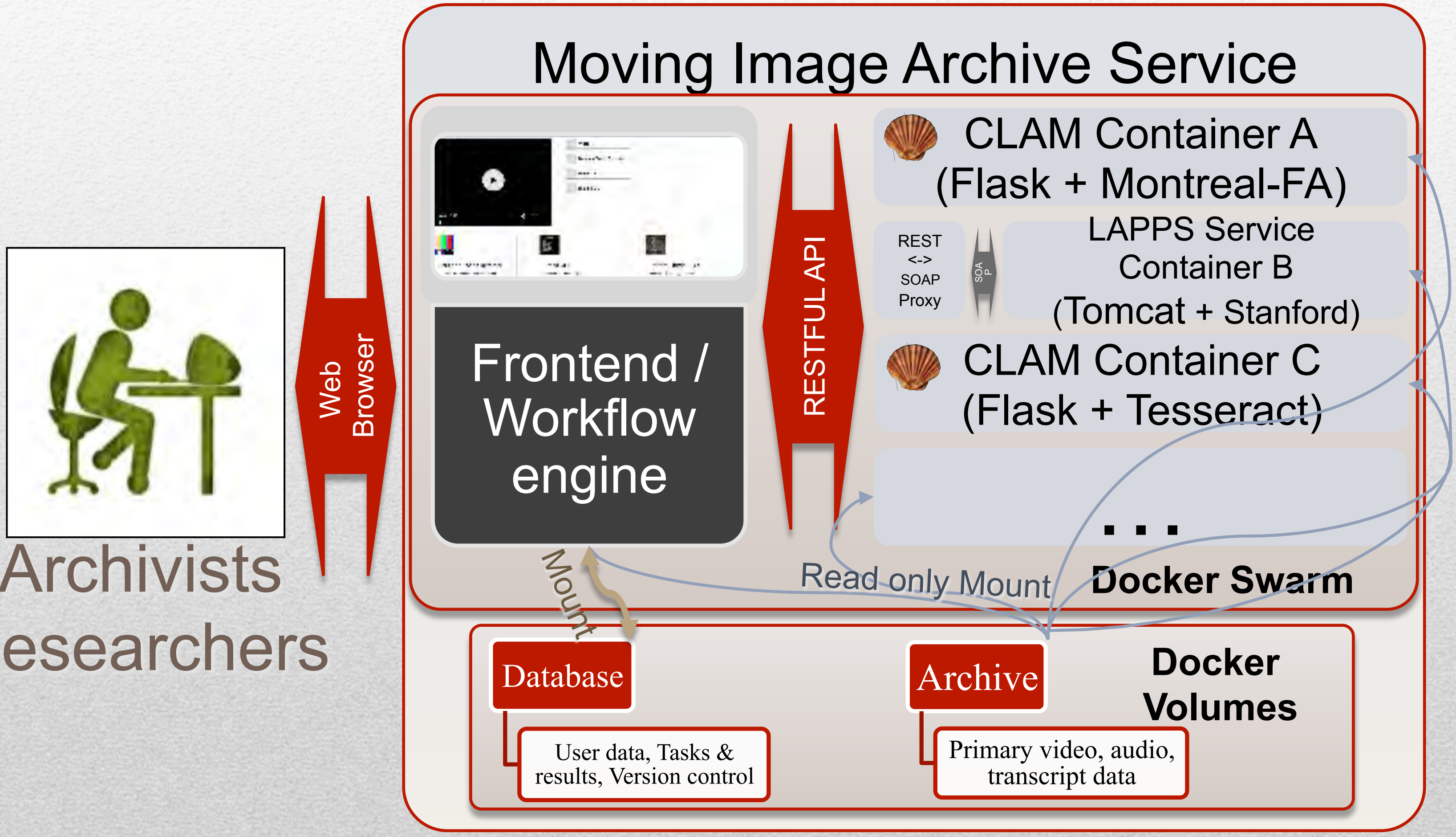


Figure 1: Integration framework

CLAMS - CL APPLICATIONS FOR MULTIMEDIA SERVICES

CLAMS Architecture



Archivists
Researchers

Web Browser

Moving Image Archive Service

Frontend / Workflow engine

RESTFUL API

CLAM Container A (Flask + Montreal-FA)

REST <-> SOAP Proxy

SOAP

LAPPS Service Container B (Tomcat + Stanford)

CLAM Container C (Flask + Tesseract)

. . .

Read only Mount

Docker Swarm

Mount

Database

User data, Tasks & results, Version control

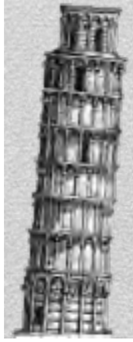
Archive

Primary video, audio, transcript data

Docker Volumes

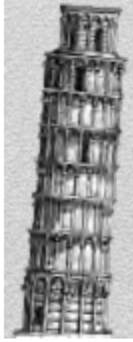
Interchange Format

- LAPPS Grid and Galaxy both are designed for text data
- All annotations in LIF always anchor on character offsets
- Successful integration with bibliographical data at HathiTrust Digital Library



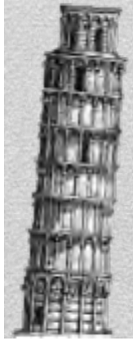
Moving towards Moving Image Archive

- Within a MIA
  - Beyond text data: Time-based media (video and audio)
  - Analysis of time-based media must be time-based, not based on character
  - BUT, MIA data are also subject to linguistic analyses!
    - Audio can be transcribed to text data
    - Video can have text inside (caption, logo, subtitle)
  - Beyond linguistic data: Need analysis tools for multimodality
    - Objects, faces, non-speech sounds, graphics, …

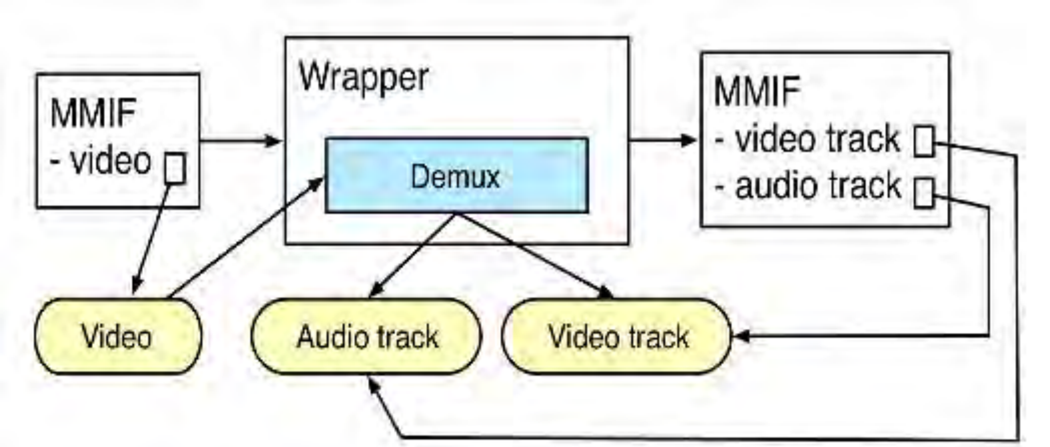


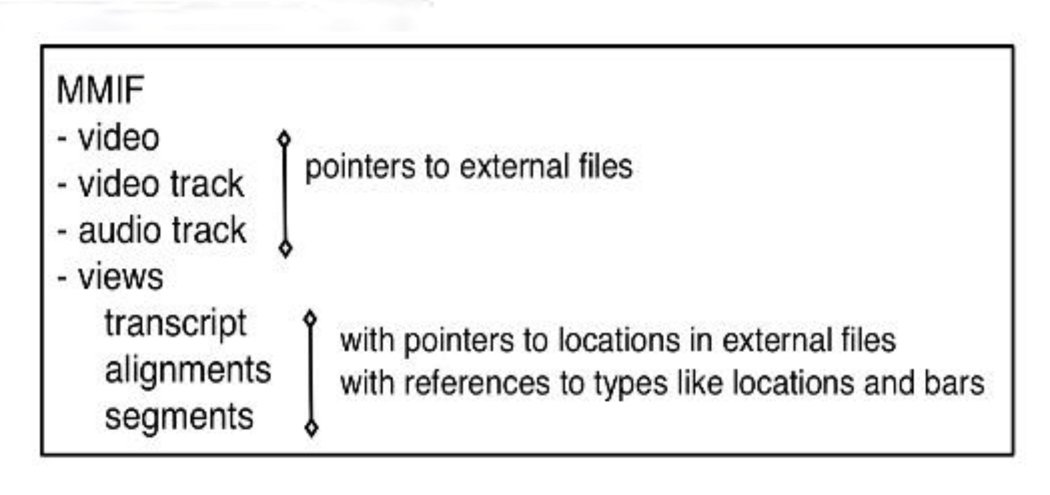
MultiMedia Interchange Format

- Handling different media types
  - Video, Audio, Text-on-image, Transcript
- Alignment between different media and annotation types is essential
  - Enables multimodal analysis
- Flexibility and Scalability of data format
  - Easy I/O, future-proof format for new technologies and tools
- Version control and Tracing
  - Video files are large in size and not easy to pass through network
  - Annotations files cannot carry primary data
  - VC and traceability are important to compensate loss in portability



Handling Primary Media




MMIF
- video
- video track
- audio track
- views

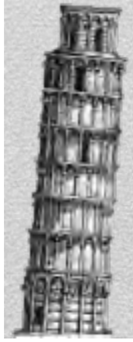
pointers to external files

transcript
alignments
segments

with pointers to locations in external files
with references to types like locations and bars

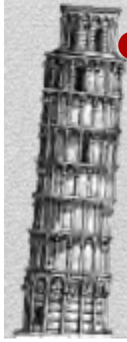
CLAMS Type Hierarchy

- Media Types
  - Time-based data – video, audio
  - Character-based data – conventional linguistic data
  - Region-base data – bounding boxes on still images
- Annotation Types bound to media types
- Special annotation types for multimodal alignment
  - E.g. forced aligner aligns time-based data and character-based data



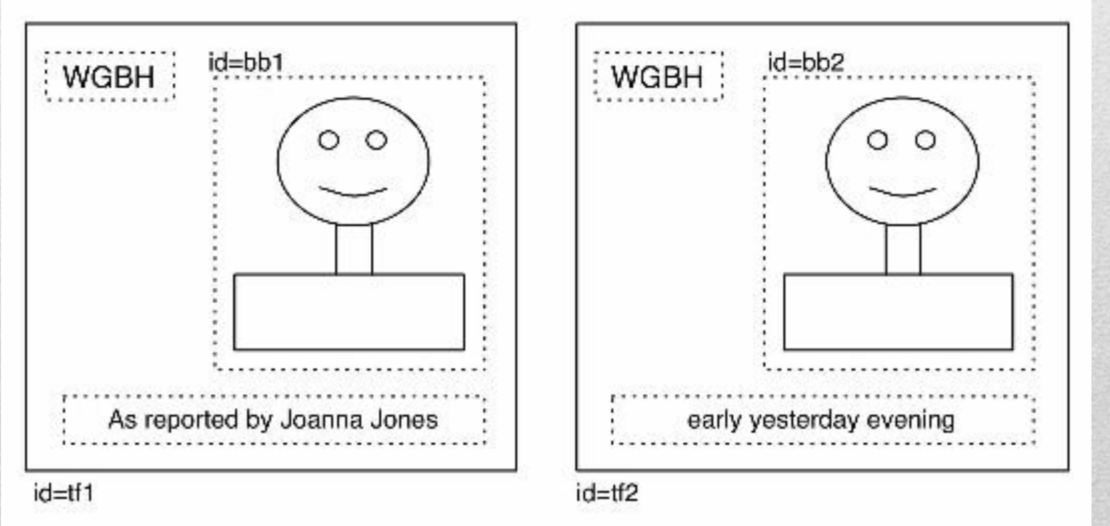
Annotation Anchors

- For text annotations: characters (start/end)
  - Named Entity, Tokenization, …
- For video or audio chunks: timestamp, frame number
  - Story Segmentation, Sound Classification, …
- For image annotations: bounding boxes
  - Face/Object Recognition, Grounding, OCR, …
- For still images from a video: bounding boxes & timestamp



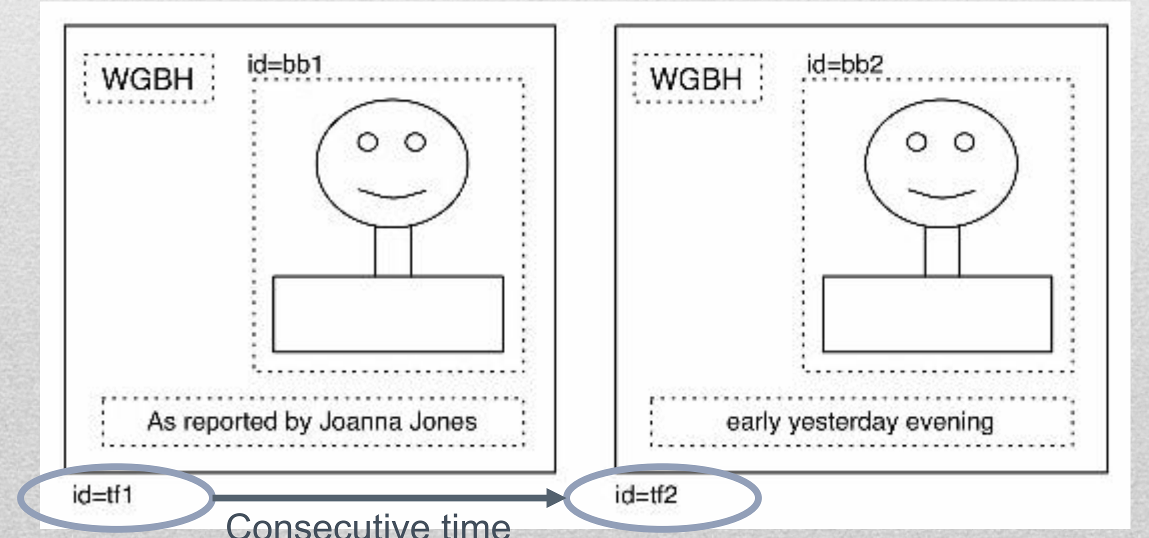
Timestamped Boxes

- What we actually want to do is to refer to some object in the video over time, for example a talking head that is on screen for some time frame
- A video object consists of a set of bounding boxes that are present in a video in a sequence of time frames. The timeframes can be consecutive or there could be gaps in case we sample one or two frames per second



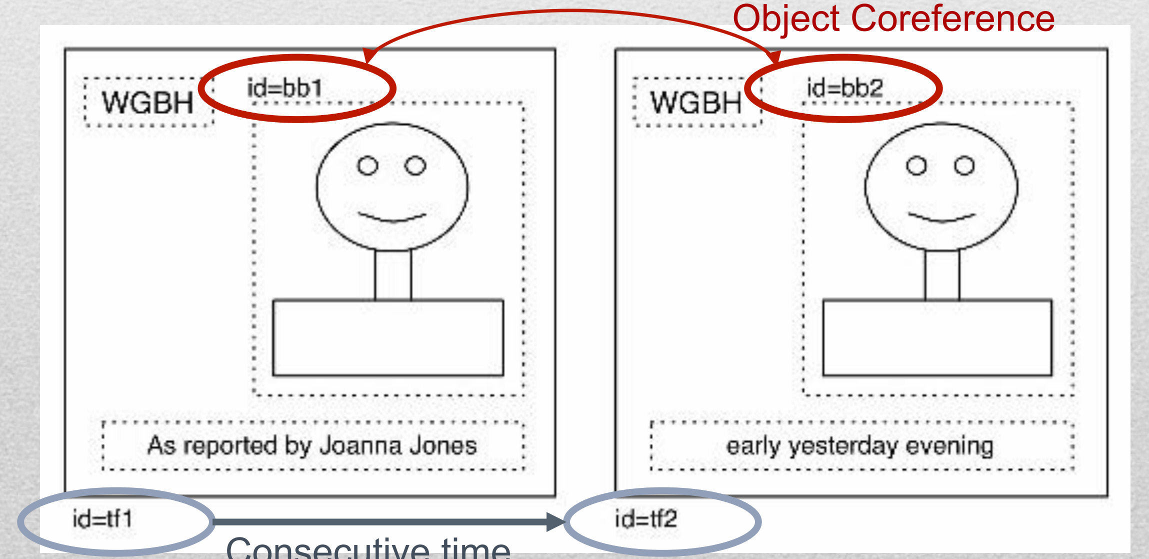
Timestamped Boxes

- What we actually want to do is to refer to some object in the video over time, for example a talking head that is on screen for some time frame
- A video object consists of a set of bounding boxes that are present in a video in a sequence of time frames. The timeframes can be consecutive or there could be gaps in case we sample one or two frames per second



Timestamped Boxes

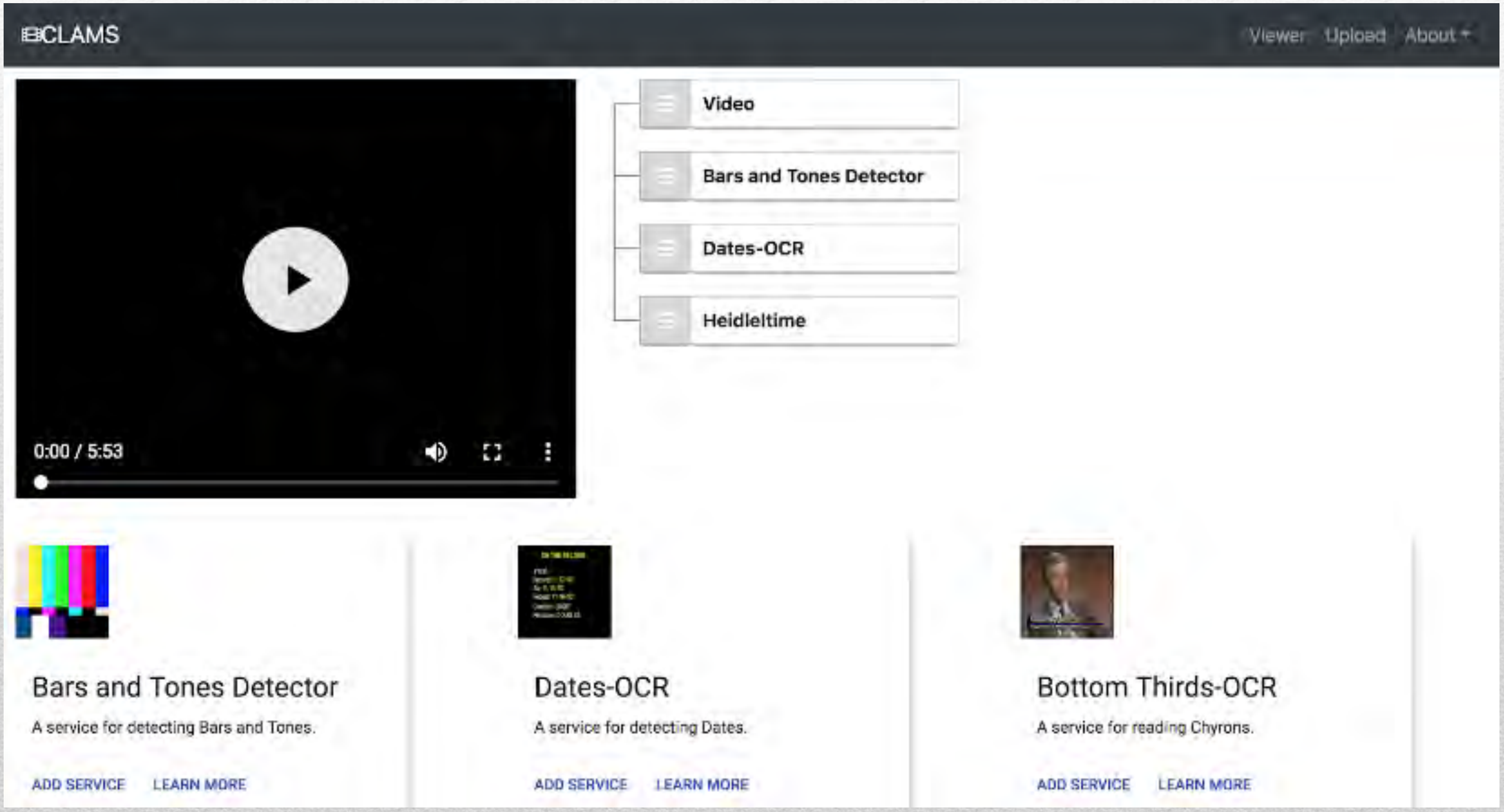
- What we actually want to do is to refer to some object in the video over time, for example a talking head that is on screen for some time frame
- A video object consists of a set of bounding boxes that are present in a video in a sequence of time frames. The timeframes can be consecutive or there could be gaps in case we sample one or two frames per second



Object Coreference

Consecutive time frame

CLAMS Platform Prototype



Bars and Tones Detector

A service for detecting Bars and Tones.

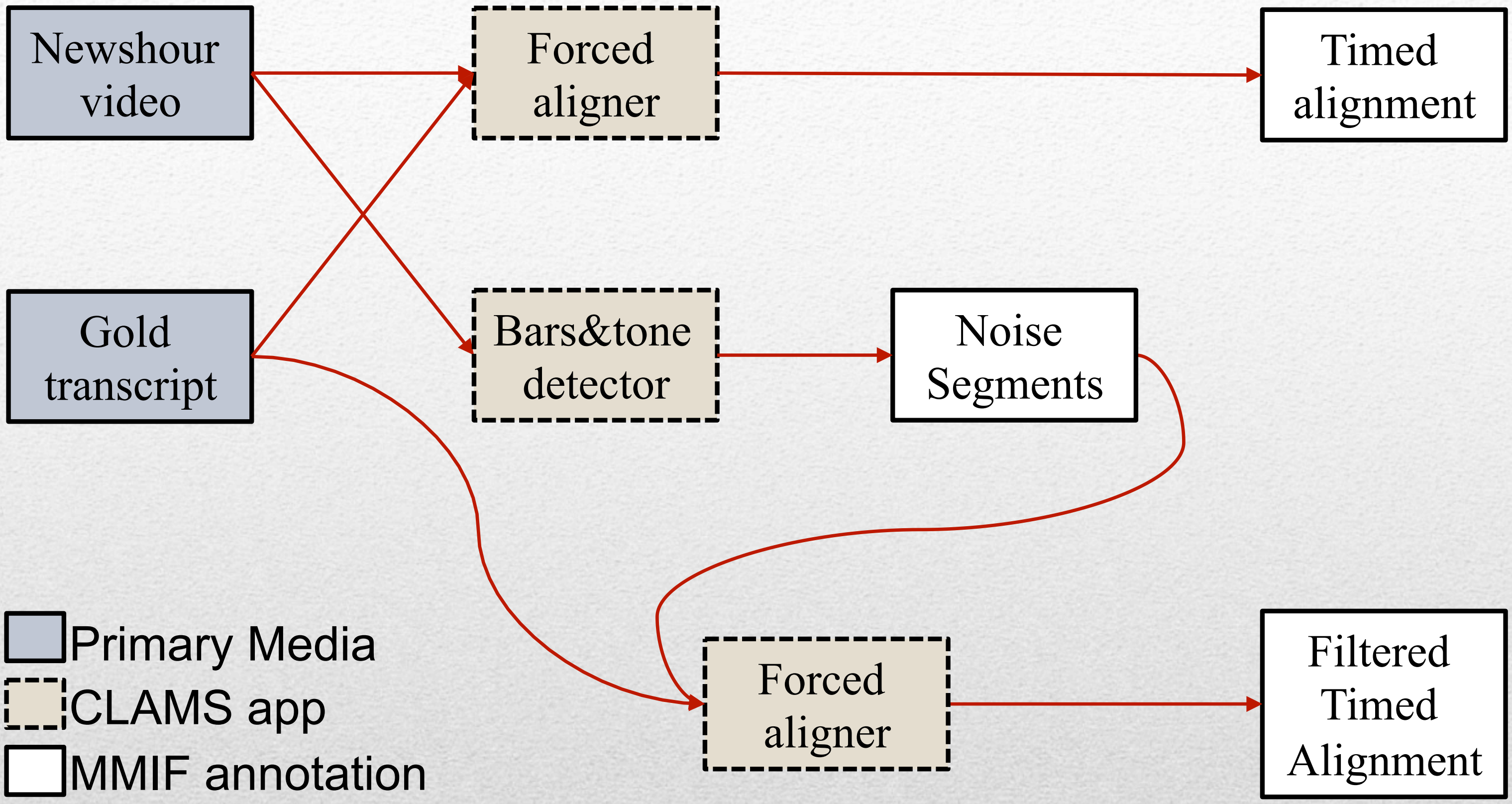
Dates-OCR

A service for detecting Dates.

Bottom Thirds-OCR

A service for reading Chyrons.

Sample workflow #1



Newshour video

Gold transcript

Forced aligner

Bars&tone detector

Timed alignment

Noise Segments

Forced aligner

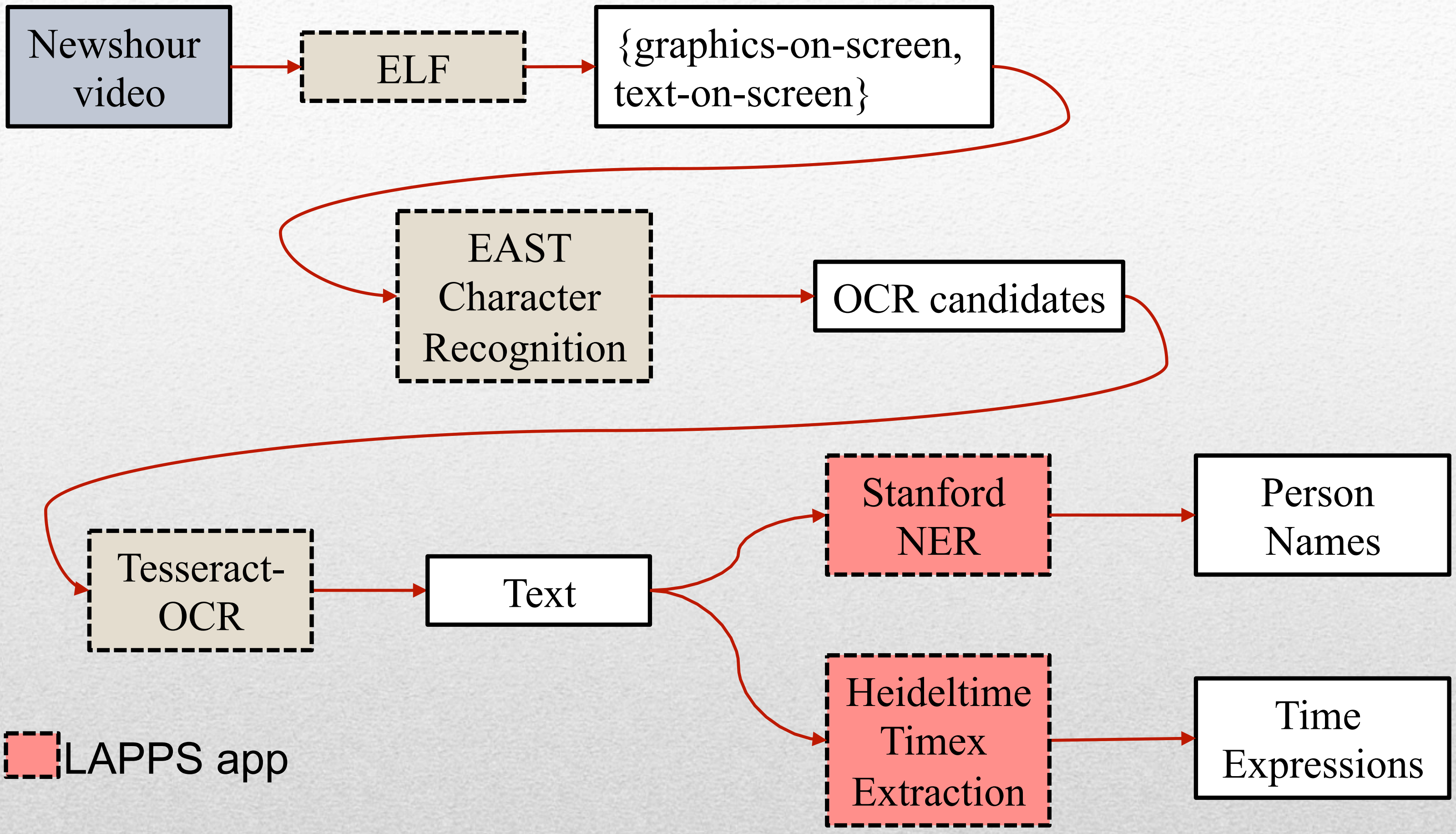
Filtered Timed Alignment

▨ Primary Media

⬚ CLAMS app

□ MMIF annotation

Sample workflow #2



Newshour video

ELF

{graphics-on-screen, text-on-screen}

EAST Character Recognition

OCR candidates

Tesseract-OCR

Text

Stanford NER

Person Names

Heideltime Timex Extraction

Time Expressions

LAPPS app

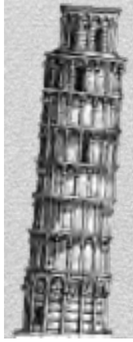
TOOL - ELF

Event Localization Finder

- what is the back story of ELF, and where did it come from?
- Question: Given a frame in a video, what is the most likely scene type?

Event Localization Finder - ELF

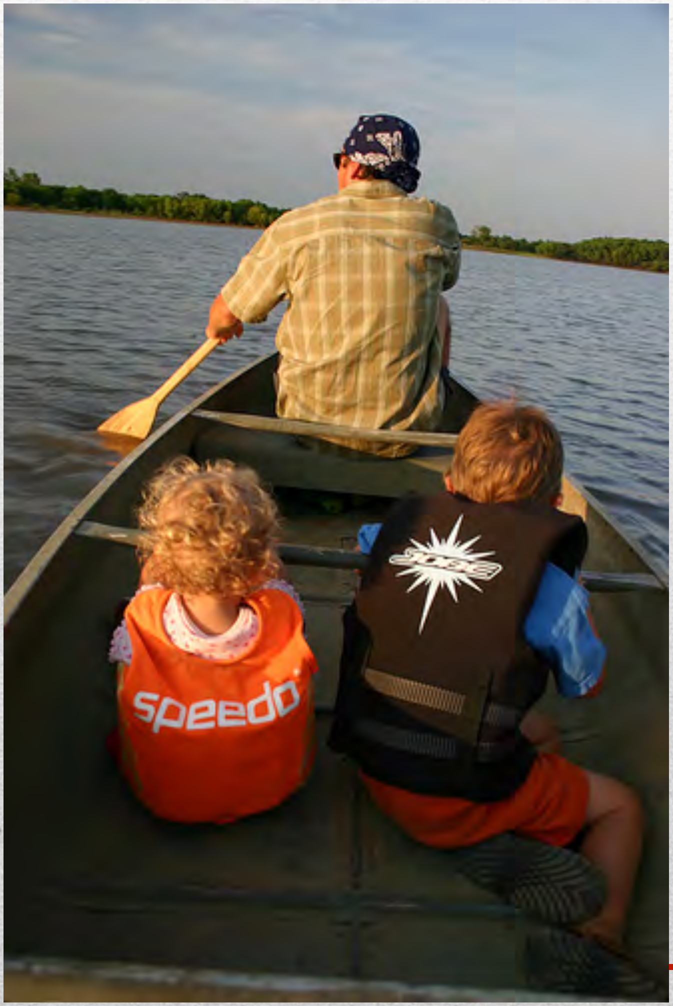
- Scene types specify where events happen
  - ○ Events and objects are contextualized by the spaces in which they are situated
  - ○ Scenes are compositions of events and objects
- Information about scene type in text is rarely explicit
  - ○ Difficult to learn presuppositions from textual data alone
  - ○ Statistical learning over multimodal corpus of events
  - ○ Mine first-order habitats and object embedding spaces



Event Localization Corpus

- An extension of the Flickr30k corpus
- Labels each image and caption with a location type
- Hierarchically structured labels similar to the Places2 database
  - allows subsumption relations among the labels

Flickr 30K Corpus



Two children wearing life jackets face an older male while he paddles the canoe they are sitting in.

A man and two children in life jackets in a boat on a lake.

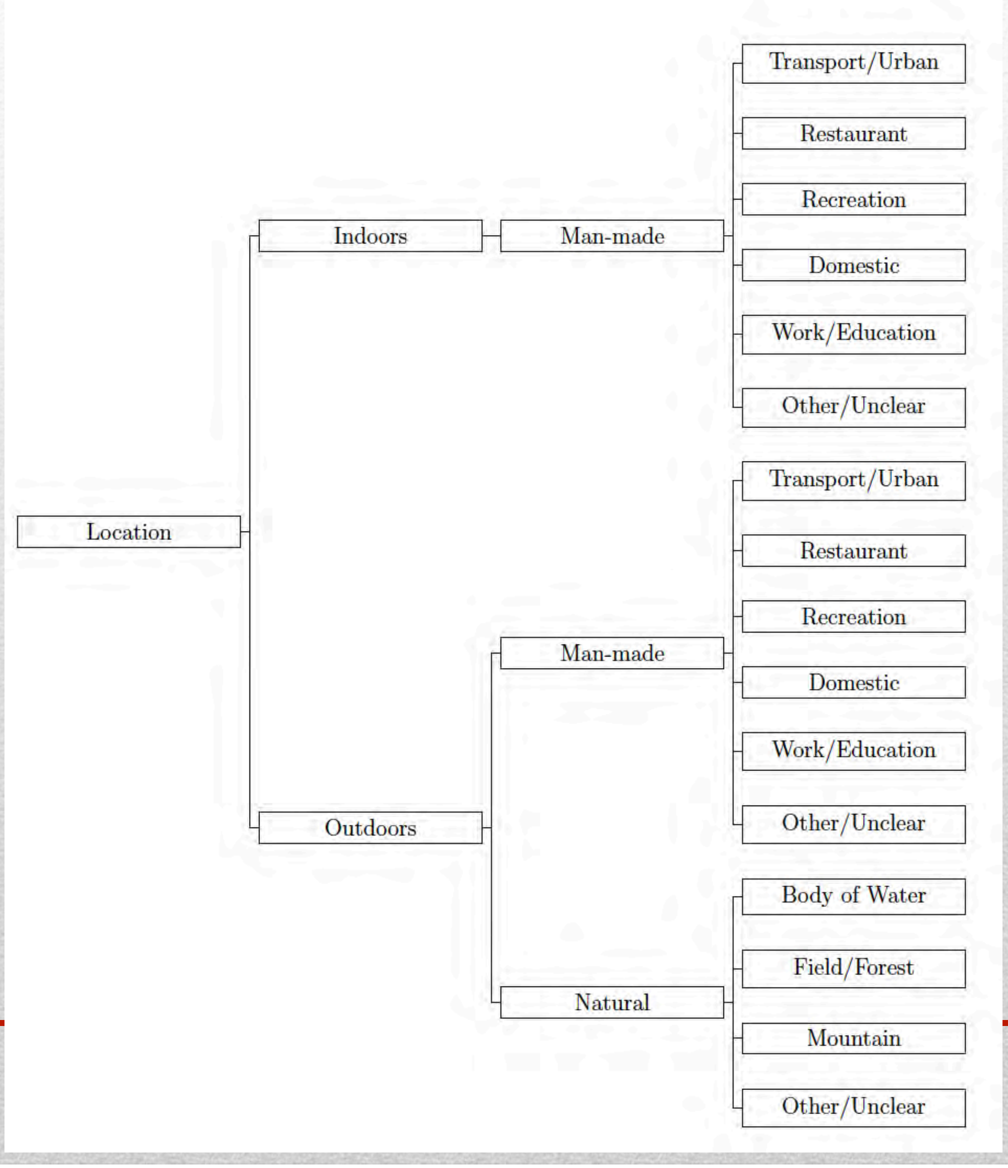
A man paddling a rowboat, with two children in back.

A man and two children in a boat on the water.

Man and two kids on a boat on the lake.

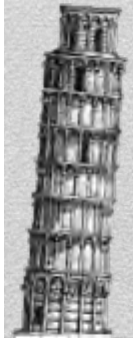
[Young et al, 2014]

Scene Type Hierarchy



Event Localization Finder

- Given an image and its caption, what is the most likely location type?
- Combines an image model and a text model
- Image model
  - Pre-trained Places2 model
    - ResNet50: 50-layer convolutional neural network
    - Extract image embedding (i.e. output of second-to-last layer)
- Text model
  - Recurrent neural network with 2 gated recurrent units



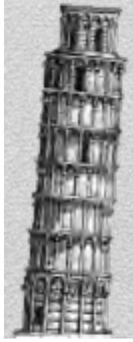
Preliminary experiments on AAPB

- Select one video (The NewsHour with Jim Lehrer, 7/22/98)
- Extract frames (1 frame/second) and annotate with scene
- Train model on annotated frames
- Select another video (The NewsHour with Jim Lehrer, 12/23/99)
- Use model to classify frames from second video

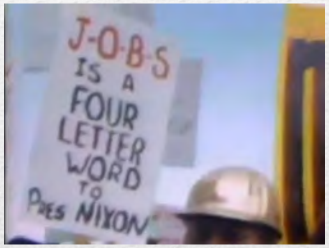
Scene ontology

- Scenes specific to TV news
  - Bars and tones
  - Blank screen
  - Graphic on screen
  - Guest in studio
  - Guest out of studio
  - Head with graphic
  - Reporter at desk
  - Text on screen

- General scene types ("in the field")
  - Indoors/man-made
  - Outdoors/man-made
  - Outdoors/natural

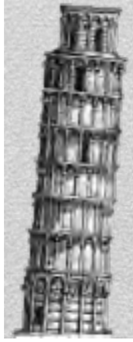


Scene ontology



Scene ontology

- Scenes specific to TV news
  - Manually annotated
- General scene types (“in the field”)
  - Automatically annotated with pre-trained model

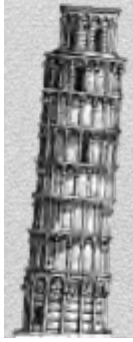


Scene ontology

- Scenes specific to TV news
  - Manually annotated
- General scene types (“in the field”)
  - Automatically annotated with pre-trained model

Model

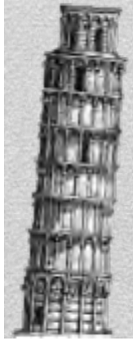
- Pre-trained model
  - Trained on Places2 database
  - ResNet50: 50-layer convolutional neural network
- Transfer learning
  - Re-train last layer to classify according to our scene ontology



Results

- Overall accuracy: 72.86%
- Results by category:

	Precision	Recall	F-measure
Bars and tones	100.00%	100.00%	100.00%
Blank screen	97.60%	97.02%	97.31%
Head with graphic	94.77%	96.67%	95.71%
Guest in studio	94.33%	75.89%	84.11%
Guest out of studio	73.40%	96.48%	83.38%
Text on screen	62.61%	95.68%	75.69%
Outdoors/man-made	77.03%	67.05%	71.69%
Indoors/man-made	63.65%	64.61%	64.13%
Outdoors/natural	83.33%	44.68%	58.17%
Graphic on screen	42.78%	36.88%	39.61%
Reporter at desk	0.00%	0.00%	0.00%



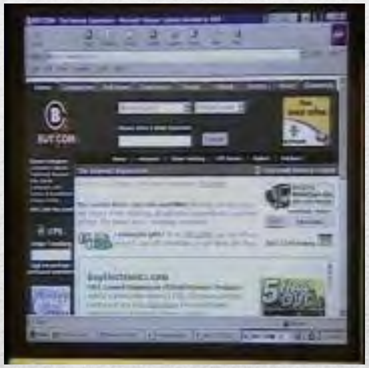
Reporter at desk vs. Guest in studio

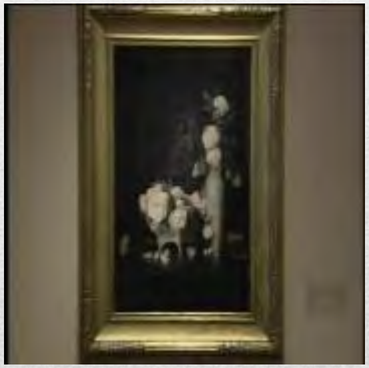
- Is this a useful distinction?

- A combined "talking head" category achieves 86.43% F-measure

Graphic on screen

- Are these frames “graphics”?



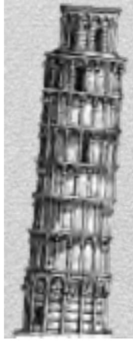


- Excluding online shopping and art segments, the graphics category achieves 86.60% F-measure

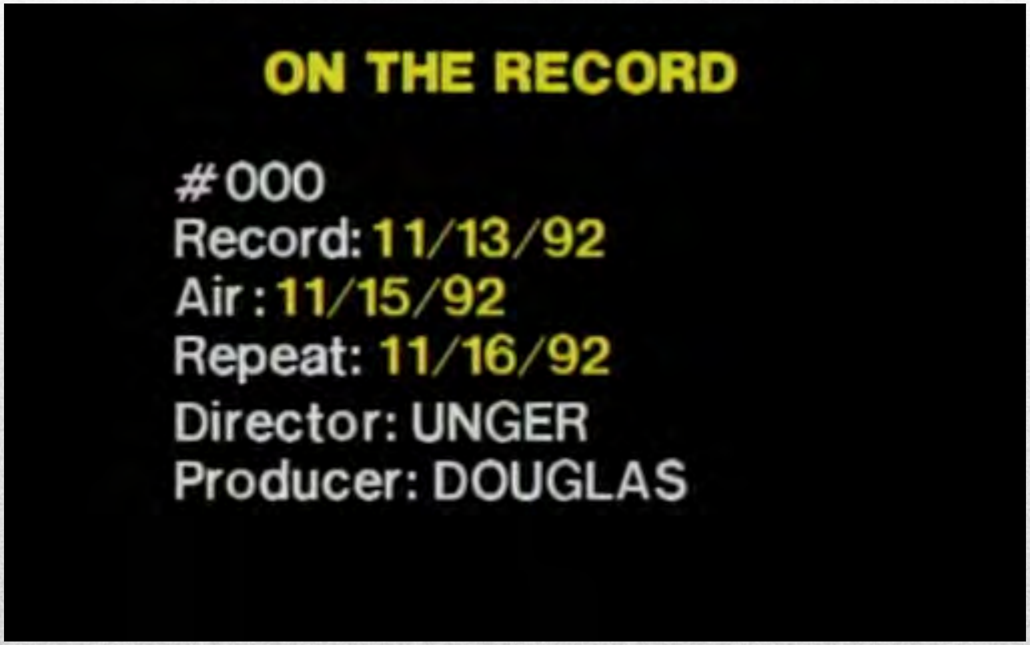
TOOL - OCR

Text Recognition within Image

- OCR on character in various background does not work well: we have to address specific tasks



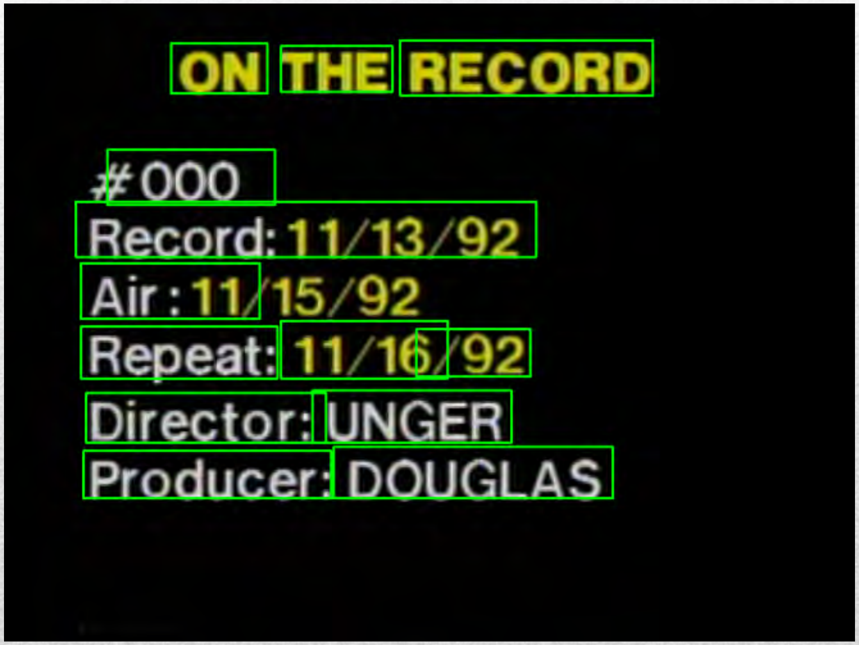
OCR Target – Digital Slate



ON THE RECORD

#000
Record: 11/13/92
Air : 11/15/92
Repeat: 11/16/92
Director: UNGER
Producer: DOUGLAS

Digital Slate – EAST Text Detection



ON THE RECORD

#000

Record: 11/13/92

Air : 11/15/92

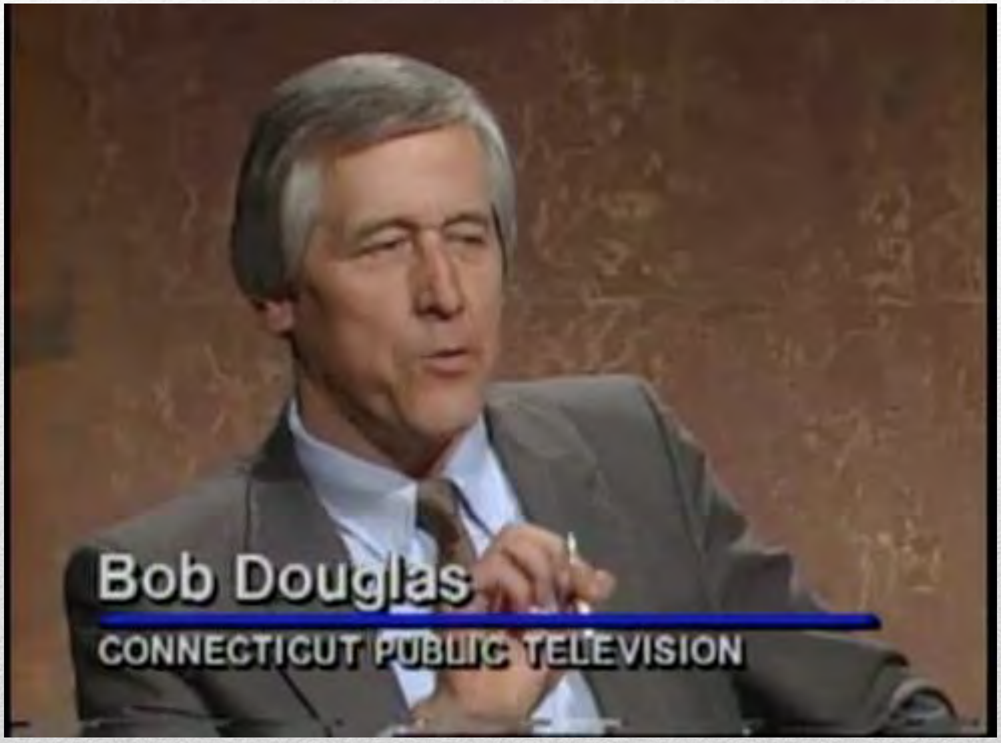
Repeat: 11/16/92

Director: UNGER

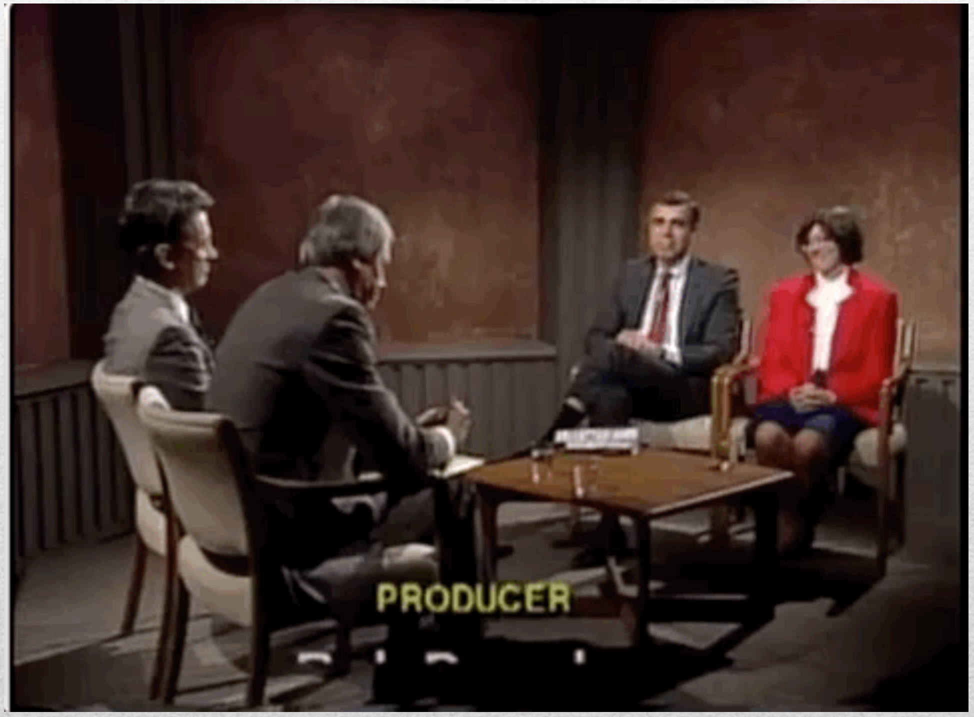
Producer: DOUGLAS

Zhou et al (2017)

OCR Target – Bottom Third



OCR Target – Credit roll



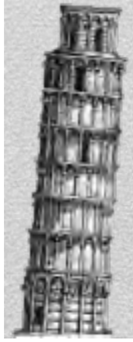
TOOL - FORCED ALIGNMENT

A Critical Bridge

- Forced Alignment is to align text transcripts to audio recordings
- This alignment is the most critical linkage for multimodal analysis – it provides methods to referencing between video and text

Limitation

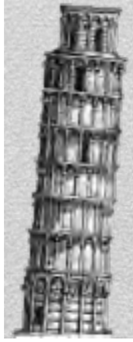
- Forced alignment technique is a byproduct of automatic speech recognition technology and going back to late 1990s
- However, as ASR is not quite perfect, FA also often suffers from poor performance – recording conditions, accents, background noise, …
- Most commonly, FA tries to align text to non-speech sounds, just like ASR tries to transcribe those



FA and CLAMS Workflow

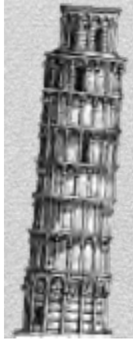
- Currently CLAMS has a wrapper tool for Montreal Forced Aligner*
- By mixing and matching different tools in CLAMS, FA can be improved as well as all downstream multimodal analysis tools!
  - Silence detection, Bar & tones detections, Non-speech classification, Tokenization, ELF

* McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*.

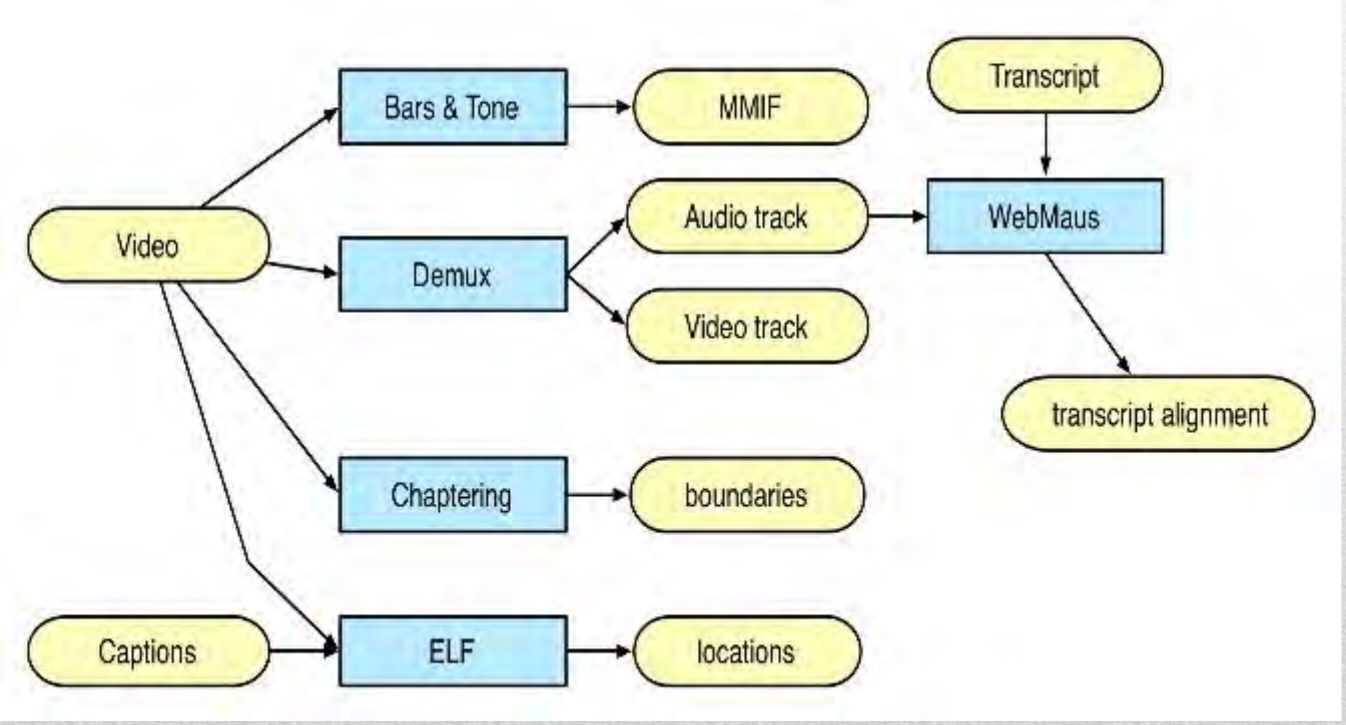


Future Work

- Chaptering – transcription-based and video-based
- Multimodal entity co-reference
- Content-based video retrieval
  - NIST has been organizing related shared tasks for 15 years, with help from BBC, Internet archive, and the Netherlands Institute for Sound and Vision
  - https://www-nlpir.nist.gov/projects/tv2018/index.html
- Full NER and parsing over transcriptions



Sample Pipeline


Transcript
Bars & Tone
MMIF
Audio track
WebMaus
Video
Demux
Video track
transcript alignment
Chaptering
boundaries
Captions
ELF
locations