Part1 - Parsing Engine
========================
Deliverables

1. Will be running as an API in backend
2. Hosted on AWS or Azure as per requirement
3. Parameters for Parsing Engine - Controlled through Admin Interface

Outputs

1. Admins datafile + metadata
2. Json file with (artist's name, exhibition title, exhibition venue, group/solo exhibition)
+ meta data. The rows which have atleast one of name, title or venue missing, will be skipped in
   JSON
3. PDF or Word of Json to compare output CV to input CV. The PDF will have entries marked red
if any of name, title or venue is missing in that row.
4. Processing time per PDF - 90 secs
5. Expected Accuracy - 80% with current scope of extracting Exhibition records.

Example (Word or PDF)
———————————————————————————————————-
ARTIST NAME - CV
DoB - <only display if a value is provided>

Solo Exhibitions
        yyyy: <title>, location

Group Exhibitions
        yyyy: <title>, location

<new page>
About your reformatted cv
Reformatted for your convenience by https://www.artbiogs.com

        About your scan   <data>
        Name: <data>
        Date of scan: <data>
        Your reference: <data>
        Your email: <data>
        Your IP: <data>
        DoB: <data>
        Scanned by artist - yes/no <data>
        For any enquiries about this scan: support@artbiogs.com

        For correcting the red marked entries open the pdf in word or use https://
www.sejda.com/pdf-editor

———————————————————————————————————

Part 2 - Website
========================

1. Would be hosted on Azure (or AWS)
2. Website would capture CV and meta data of scan and call backend API to parse CV
3. It would display error message if any
4. The original and parsed CV would be displayed for comparison
5. Metadata includes the name of the artist, email of the person scanning the cv, a unique id for the scan, data and time of the scan and a couple of Booleans tba
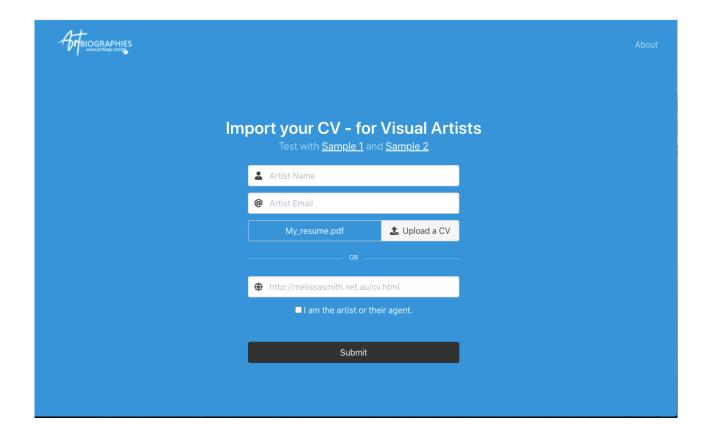
Administrator
1. Web form listing all scans with hyperlinks to the INPUT doc, the OUTPUT doc and the json file along with a value to indicate that the cv was successfully parsed. This value will come from the webform after the scan and be input by the user.
2. The Administrator's webform will be populated from a small database or file
3. Contain all the parameters that change the behaviour of the parsing engine

Outputs
1. A json file for each cv
2. A word/pdf document that enables the user to compare the input to output in an easy manner on the website

Web Form Screenshot
========================

John Doe

Born 1973
Sydney

Selected Solo Exhibitions

2017	*Take Me Home*, Michael Reid Sydney
2015	*4 Solos: Marc Etherington*, Hazelhurst Regional Gallery, Sydney
2014	*Marc Making*, Gallery 9, Sydney
	Marc Etherington, Breezeblock Gallery, Sydney
2013	*Two Sydney Painters* (with Digby Webster), D.N.A. Gallery Sydney
2012	*Life and Times*, Firstdraft Gallery, Sydney

Selected Group Exhibitions

2017	Michael Reid at Sydney Contemporary 2015 (Art Fair), Sydney
	You're my no. 1: curated by Talia Smith, Firstdraft, Sydney
2016	*Seven deadly sins*, Old Rum Store, Chippendale
	A Billion bucks, The Honeymoon Suite, Brunswick, Melbourne
2015	Michael Reid at Sydney Contemporary 2015 (Art Fair), Sydney
	Here, there and everywhere, Sheffer Gallery, Sydney
	Biggie Smalls, Casula Powerhouse Arts Centre, Sydney
	Group Show, Gallery 9, Sydney
	Friends With Benefits, Twenty Thirty Seven, Sydney
2014	*Is it a bird? Is it a plane?* Hawkesbury Regional Gallery, Sydney
	Glazed and Confused, Hazelhurst Regional Gallery, Sydney
	One Last Time (curated by Peter Fay), Sheffer Gallery, Sydney
	Still Nothing New (curated by Sophie Kitson), William St Windows, Sydney
	John Fries Art Award, UNSW Galleries, Sydney
	Fisher's Ghost Art Award, Campbelltown Arts Centre, Sydney
	Renegades, KickArts Contemporary Arts, Cairns (touring)
2013	Art On Paper Award, Hazelhurst Regional Gallery, Sydney
	Salon des Refusés, S. H. Ervin Gallery, Sydney
	Home and Hosed, Hazelhurst Regional Gallery, Sydney
	Exit Pursued by a Bear, Sheffer Gallery, Sydney
2012	*Out of the Blue*, Sheffer Gallery, Sydney
2011	*Showdown*, Sheffer Gallery, Sydney

EXTRACTION DETAILS

Reference ID	7391dda5dbfd88da6b51b3dff6cecdc6
Name	
DOB	
Solo Exhibitions	
Group Exhibitions	

FILE DETAILS

CV	
Textract	
Parsed JSON	
Parsed PDF	

Admin Page Screenshot
=======================

Example CV Output
=====================

Kent Morris - CV
Parsed with https://artbiogs.com
Name Kent Morris DOB 1996

Solo Exhibitions:
2019: Barkindji Blue Sky , Vivien Anderson Gallery, Melbourne VIC
2019: Unvanished , Horsham Regional Art Gallery, Horsham VIC
2019: Unvanished , Kluge-Ruhe Aboriginal Art Collection, University of Virginia USA 2017:
Unvanished , Vivien Anderson Gallery, Melbourne VIC
1995: My Life as Daryl Hannah , Centre for Contemporary Photography, Melbourne 1993:
Portraits of Pleasure , Luba Bilu Gallery, Melbourne
1991: Primary Narcissism , Luba Bilu Gallery, Melbourne
1990: Double Infant , Kent Fine Art, Melbourne
1989: Kitchen Child , Kent Fine Art, Melboume
1988: Capital , Christine Abrahams Gallery, Melboume

Group Exhibitions:
2019: 36th Telstra National Aboriginal & Torres Strait Islander Art Award , MAGNT, Darwin NT
2019: Celebrating Culture: Contemporary Indigenous Art , Glen Eira Gallery, Melbourne VIC 2019:
Echo Chambers: Art and Endless Reflections , Deakin University, Melbourne VIC
2018: Afterimage , Deakin University Art Gallery, Melbourne VIC
2018: 35th Telstra National Aboriginal & Torres Strait Islander Art Award , MAGNT, Darwin NT 2018
2018: King and Wood Mallesons Contemporary ATS Art Prize , Parliament House, Sydney NSW

2018: Wyndham Art Prize , Wyndham Art Gallery, Melbourne VIC
2018: Tell , Sydney Festival, UNSW Galleries, Sydney NSW
2018: Foundation , Fairfield City Museum and Gallery, Sydney NSW
2017: Tarnanthi , Art Gallery of South Australia, Adelaide SA
2017: Darebin Art Prize , Bundoora Art Centre, Melbourne VIC
2017: Bowness Photography Prize , Monash Gallery of Art, Melbourne VIC
2017: 34th Telstra National Aboriginal & Torres Strait Islander Art Award , MAGNT, Darwin 2017: Tell , Ballarat International Foto Biennale, Ballarat VIC
2017: Sunshine Coast Art Prize , Caloundra Regional Gallery, Caloundra QLD 2017: Footscray Art Prize , Footscray Community Art Centre, Melbourne VIC 2016: Sovereignty , Australian Centre for Contemporary Art, Melbourne VIC 2016: The Incinerator Art Award , Incinerator Gallery, Melbourne VIC
2016: 33rd Telstra National Aboriginal & Torres Strait Islander Art Award , MAGNT, Darwin NT
2016: Reframed , Incinerator Gallery, Melbourne VIC
2016: Wyndham Art Prize , Wyndham Art Gallery, Melbourne VIC
2015: Darebin Art Prize , Bundoora Art Centre, Melbourne VIC
2015: The Incinerator Art Award , Incinerator Gallery, Melbourne VIC
2015: Victorian Indigenous Art Awards , Art Gallery of Ballarat, Ballarat VIC 1995: Photo Posters , Centre for Contemporary Photography, Melbourne
1993: The Margaret Stewart Collection , National Gallery of Victoria, Melbourne 1989: Proposals , George Paton Gallery, Melbourne
1988: Site of Execution , Australian Centre for Contemporary Art, Melbourne 1988: Artisans , George Paton Gallery, Melbourne

Proposed Solution
1. CV Parser Pipeline (Python Code)
    A. Code to apply Textract model to PDF to extract the text from PDF.
    B. Identification of Section Headings using string match with gold list of heading words. The section identified would be the exhibitions.
    C. Manual Tagging/ Labelling of extracted texts to prepare the data for Training the NER model. The labels would be <Year>/<year range>, <exhibition title>, <exhibition venue>, <exhibition city>
    D. The model will then be able to parse and tag any paragraph with these tags, automatically
    E. Once tags are extracted, an algorithm to map these values together and display the required output
    F. Customisation on top of this would be implemented to handle special cases.

2. The Inference pipeline would be built in the backend Azure server in Flask and exposed as a parser API (Flask API). This API would accept the PDF data from the front end Website Application and apply the stored model hosted in Azure

3. The website front end would be developed in React JS and hosted on Azure, along with the backend which would be a Flask or Java service.

4. The admin front end would also be developed in Reach JS and hosted on Azure, and in backend would be a Flask/Java service that would connect with the databases of parsed resumes.

Architecture Diagram