National-AI Enabled Repository for Wales

Evans, Gemma; Thomas, Cory; Sauze, Colin; Zwiggelaar, Reyer; Higgins, Sarah

DOI:
10.20391/z11x-tw26

Publication date:
2022

Citation for published version (APA):
Evans, G., Thomas, C., Sauze, C., Zwiggelaar, R., & Higgins, S. (2022). *National-AI Enabled Repository for Wales: AI Prototype Testing Sessions Report*. Prifysgol Aberystwyth | Aberystwyth University. https://doi.org/10.20391/z11x-tw26

National-AI Enabled Repository for Wales

AI Prototype Testing Sessions Report

October 2022

Authored by Gemma Evans (PDRA, Aberystwyth University)

Aberystwyth University Project Team: Reyer Zwiggelaar (Principal Investigator); Sarah Higgins (Co-Investigator); Cory Thomas (Software Engineer); Colin Sauze (Software Engineer); Vicky Jones (Project Administrator); Gemma Evans (PDRA).

Project Partners: National Library of Wales; Royal Commission on The Ancient and Historical Monuments of Wales; Archives and Records Council Wales; Cadw; Canolfan Bedwyr; Geiriadur Prifysgol Cymru; Digital Preservation Coalition; Eisteddfod Genedlaethol Cymru; Wales Higher Education Libraries Forum; and Welsh Government.

Funded by AHRC National Infrastructure for Digital Innovation and Curation in Arts and Humanities (iDAH) Grant AH/W007487/1

Contents

1. Executive Summary

This report provides an overview of the AI prototype testing sessions carried out as part of the National AI-Enabled Repository for Wales Project which formed part of the initial scoping grants stage of the AHRC's National Infrastructure for Digital Innovation and Curation in Arts and Humanities (iDAH). It outlines how the activities in each session were carried out, the results and feedback generated, and key findings and recommendations for future testing.

Two testing sessions were conducted in September 2022 to analyse how the outputs of our AI-Enabled Repository prototype compared against human-generated responses, with a focus on the machine-learning based functions of image analysis, text summarisation, and ontological text classification. The first session took place during Aberystwyth University's AI Research Hub Symposium (*Pushing Humanity Forward: Our AI-Enabled World*) and participants included symposium attendees from a range of academic disciplines. The second was conducted online and was attended chiefly by participants of focus groups held previously during the project including research data managers, GLAM (Galleries, Museums, Libraries and Archives) specialists, and academic researchers. The differences in the professional backgrounds and assumed knowledge bases of participants across the two sessions offered an opportunity to understand how different types of project stakeholder might engage with AI testing activities. The sessions also allowed us to examine the extent to which such activities can generate discussion and engagement, facilitate a better understanding of the functions of a National AI-Enabled Repository, and inform both the development of our AI prototype and the design of further testing activities in the future.

2. Key findings and recommendations for future testing

2.1 Key findings

2.2.1 **Outcomes:**

- Successful demonstration of the feasibility of carrying out group testing activities in-person and online
- Successful inclusion of a range of participants, with differing professional backgrounds and AI experience levels, in interactive activities and discussion
- Demonstration of testing activities for a range of machine-learning based AI outputs including image segmentation and labelling, text summarisation, and ontological text classification.

2.2.2 **Findings**

- Among the image analysis tests, the semantic segmentation model generally generated fewer results than the image detection model, but with a higher degree of overlap with human-generated responses.
- In grading the prototype's outputs, participants scored the semantic segmentation model more highly than the image detection model for image analysis activity (Activity 1), but gave them the same accuracy rating for the second image analysis activity (Activity 2).
- The number of activities which participants provided accuracy ratings for was too low to draw broad conclusions as to the performance of the prototype's image analysis

functions, but the feedback on the grading activities provided useful input for the design of future testing sessions (see section 2.2 below).

- Participants in session 2 provided a lower number of classifications for both the image analysis activities and the ontological text classification task. The level of technical detail in the classifications provided by the online participant groups was also higher. This may have been influenced by the backgrounds of the participants in each session, but may also reflect differences in the setup of in-person versus online activities
- Participants in session 1 raised some concerns over the fairness of giving accuracy ratings for the AI outputs without access to information about how the model was trained, including classification datasets
- Given that only four activities were carried out within a two-hour timeframe and participants suggested they would not want the session to be longer, generation of a larger number of results from future testing may require:
  - Running more testing sessions; and/or
  - Running some activities on an individual rather than group basis.

2.2 Recommendations for future AI testing

2.2.1 The following are recommendations drawn from our two testing sessions for consideration in planning future AI testing activities as part of a broader National AI-Enabled Repository for Wales project:

- Once the existing functions of image and text analysis have been trained on a wider range of input data, the testing activities could be repeated to monitor the impact on accuracy of the outputs
- Testing activities should be developed for the types of input data which have not yet been tested including:
  - Audio and audio-visual materials
  - Geo-spatial data
- Testing activities should be designed for assessment of additional AI functions including:
  - Bilingual translation
  - Optical character recognition (OCR)
  - Linked data functions
- In planning testing activities for new types of data and AI functions, some level of expert or subject-domain knowledge may be required including:
  - Palaeography skills for OCR-based tasks, especially for older handwritten texts
  - Fluency in the Welsh and English for bilingual interpretation functions and Welsh language AI outputs
- The classification datasets the AI model is trained on for tasks such as image analysis and ontological text labelling should be provided to participants for review when grading the AI model outputs
- Future testing activities which aim to test the accuracy of the model may require more specific guidance for participants, including whether the absence of anticipated classifications should be considered in addition to the accuracy of classifications actually applied
- It may be useful to maintain an element of open-choice responses, however, as a first step in any testing activities as this can help to highlight the diversity of participants' interpretations and indicate any limitations in the classification datasets themselves.

3. Overview of Testing Sessions

3.1 AI Prototype functions

3.1.1 Functions tested

Figure 1 below provides an overview of the functions of the AI-Enabled repository prototype. These functions include data conversion; machine learning applications including audio, text, and image analysis; metadata generation; metadata linking; and bilingual translation. The processes of text and image analysis, highlighted in the diagram with a circle, formed the focus of our testing activities. The specific AI functions tested included:

- Image analysis, including:
  - Image detection
  - Semantic segmentation
- Text analysis, including:
  - Text summarisation
  - Ontological classification ("importance labelling")

These functions were selected because text and image files formed the bulk of the datasets which the AI prototype was trained on and because text and image sources were easier to use in group testing activities than other data types (e.g. audio). Image analysis, text summarisation and classification are also elements of NLP-based machine learning that were considered the most appropriate for use in our testing sessions as they can be more effectively compared against human responses without the need for expert knowledge.

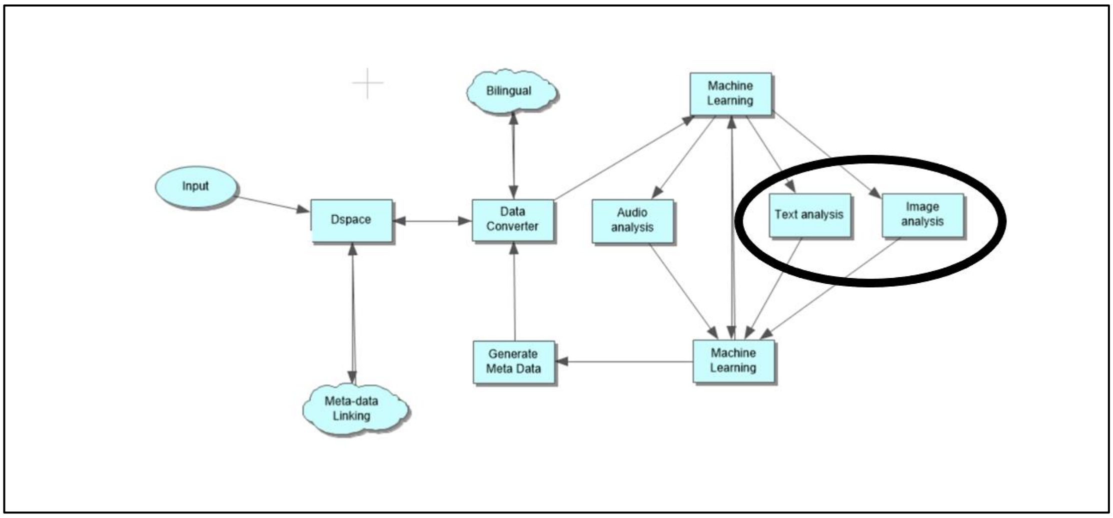
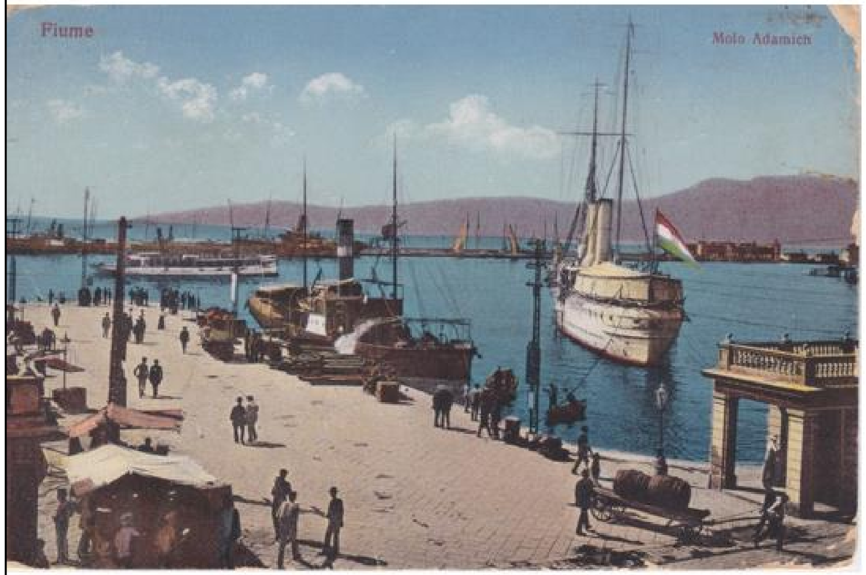


Figure 1: Prototype overview

3.2 Focus group 4 testing exercise

3.2.1 The two dedicated testing sessions run in September 2022 built on an exercise carried out during our fourth focus group on the theme of Knowledge Discovery which took place online on 14th July 2022. Participants were asked to work in groups in breakout sessions to provide labels for an image (Figure 2 below). They were allowed to select from an unlimited range of words and were not restricted to the same classifications that the AI semantic segmentation and image detection models were trained on. Participants' results were compared with those generated by the AI prototype during a discussion in the round. The results of the activity suggested that while the AI model produced a limited number of results, those which it did produce broadly matched some of those selected by focus group participants. In addition, the activity and the group discussion which followed it suggested that the "ground truth" surrounding the image's contents was subjective, as 41 out of the 54 terms suggested by humans were unique, and indicated there may be a variation in approaches if people are given freedom to select from an unlimited range of words to classify the objects within an image.

Image



AI prototype results

Image detection model:

- Dock, dockage, docking facility
- People, group of people

Semantic model:

- Sea

Participant responses

boat, steam power, wind power, people, hills, cart, barrel, dock, timber, awning, sea, water, flag, balustrade, lamp post crates, clouds, quay, paving, column, wheel, funnel, crane, ropes, moorings, rowing boat, blue, red, white, green

Steam ship, Harbor, Trade, Transport, Boats, Sea, Rigging, People, Costume, 18 century, Quay, Barrel, Croatia Flag

columns, maritime, historic photograph, tinted postcard, trade, sea, sailing, barrels, port, industry, Italy, fiume, port, steam, harbour, boats

port, harbour, steamboat, cargo, mountains, warehouses, rigging, flag, people, transport, trade, funnels, sails, sailors, barrels

Figure 2: Focus group 4 Testing Exercise Results

3.3 Testing Session 1: AI-Enabled Research Hub Symposium

3.3.1 **Session 1** took place as part of Aberystwyth University's AI Research Hub Symposium (*Our AI-Enabled World: Pushing Humanity Forward)* on 23rd September 2022. The event was held in-person and participants included symposium presenters and attendees, the majority of whom were Aberystwyth University staff or post-graduate research students. Attendees included researchers with a background in computer science, as well as those working on AI-related projects from a broader range of disciplines including information studies, business, life sciences, psychology, and physics. Topics addressed by other presenters in the symposium ranged from bioinformatics to AI risk management and its applications for psychological health interventions. The AI-Enabled Repository Testing Activities session followed on from a presentation given by project members on the theme of 'Developing a National AI-Enabled Repository for Wales which introduced participants to the broader context of the AI prototype's development and functions.

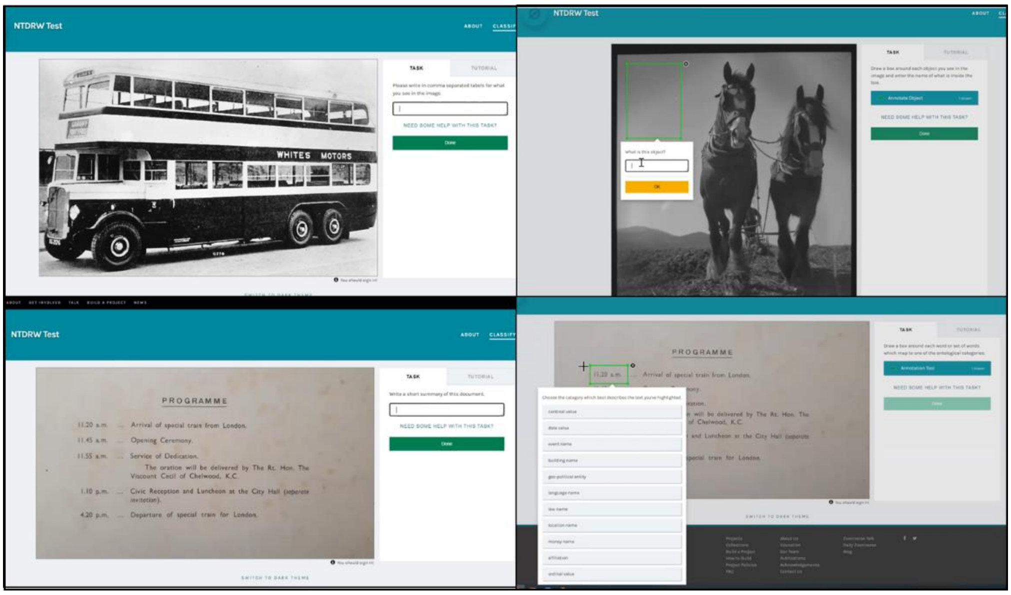
3.3.2 **Event structure:** Attendees of the symposium who participated in the testing session were divided into four groups of between three and five people. They were provided with an introduction which outlined the types of activities involved and the ethical considerations we had taken into account when planning the session (see section 3.5 below). Each group of participants, but for the majority of the tasks they used pens and paper. After each task participants were invited to share their answers and compare them with the AI model's responses. An opportunity to provide feedback was also provided at the end of each task and towards the closing of the session. Their comments have been included in the relevant activity sections below.

3.4 Testing Session 2: Online event

3.4.1 **Session 2** took place online via Zoom on Tuesday 27th September. The session included 19 participants. A large majority of them had previously attended one or more of our project focus group sessions. They included academic researchers (humanities, law, information studies, and computer science), research data managers, archive managers, digital preservation specialists, librarians, and individuals with expertise in archaeological heritage and image interoperability. The institutions represented comprised:

- Aberystwyth University
- National Library of Wales
- Royal Commission on the Ancient and Historical Monuments of Wales
- Cardiff University Special Collections
- Peoples Collection Wales
- Digital Preservation Coalition
- Conwy Archives
- Pembrokeshire Archives
- International Image Interoperability Framework (IIIF)

3.4.2 **Event structure:** A brief presentation on the project including an overview of previous focus groups and our AI prototype functionality was provided at the beginning of the session. Participants were divided into two breakout groups of around 7-8 people for each activity. To carry out the activities online, project software engineer Colin Sauze created an online activities environment within Zooniverse which participants used to view, tag and annotate images and text sources, as well as to draw bounding boxes for image and text segmentation tasks (see Figure 3).[1] Following each activity, there was an opportunity to share each group's answers and compare them with the AI model's responses. As in the symposium session, the online event also provided an opportunity for participants to give feedback on each activity and on the session generally and their comments have been included in discussion of the relevant activity below.



Please write in comma separated labels for what you see in the image.

Write a short summary of this document.

Draw a box around each object you see in the image and enter the name of what is inside the box.

Draw a box around each word or set of words which map to one of the ontological categories.

Choose the category which best describes the text you've highlighted.

PROGRAMME

11.20 a.m. ... Arrival of special train from London.
11.45 a.m. ... Opening Ceremony.
11.55 a.m. ... Service of Dedication.
The oration will be delivered by The Rt. Hon. The Viscount Cecil of Chelwood, K.C.
1.10 p.m. ... Civic Reception and Luncheon at the City Hall (separate invitation).
4.20 p.m. ... Departure of special train for London.

Figure 3: Zooniverse online activities environment

3.5 Ethics

3.5.1 Prior to conducting the testing sessions with participants, including the testing exercise undertaken as part of our knowledge discovery focus group, an ethical assessment was undertaken and approved by Aberystwyth University.[2] A version of the following statement was read to participants prior to the beginning each testing session and was replicated on the Zooniverse testing pages.

3.5.2 **Ethics statement:** "The research exercise these images will be used in involves asking participants to describe a set of archival images for the purposes of comparison against an AI prototype's analysis of the same. Participants will also be requested to provide feedback on the accuracy and quality of the results provided by the AI model. Participants will not be required to provide any personal data (including special category data) as part of the exercise. Participants will be requested to work together in groups and submit their answers collectively and these will be anonymously recorded. The images used in this activity have been pre-screened to ensure that they do not contain the personal data of any living individuals and that no living individuals can be identified within the images. The images have also been assessed to ensure that they are not reasonably likely to cause offence, harm or distress to

[1] https://www.zooniverse.org/projects/colinsauze.ntdrw-test

[2] Approved application reference: 23145

participants. The intended use of these images as part of this exercise has been approved in an ethics assessment conducted through Aberystwyth University."

4. Activity 1: Image analysis exercise (Image 1)

4.1 Activity design

4.1.1 **Session 1 activity design**: Each group was provided with a printed copy of a black and white photographic image of a six-wheeled bus (Image 1) and were requested to label it with as many words as they felt appropriate to categorise its contents.[3] Following the labelling activity, participants were shown the results of the AI prototype and were asked to grade its results using a 10%-100% scale in 10% intervals. They entered their results via the online app Slido.

4.1.2 **Session 2 activity design**: Each breakout group was requested to view the Zooniverse webpage which included a copy of the original image and a text box to write labels to categorise the image. Following the labelling activity, participants compared their results to the AI prototype outputs and provided a grade for the model using a 0%-100% in 20% intervals. Due to the time constraints of the session, participants did not re-enter the breakout rooms to provide their grades and instead submitted them on an individual basis.

4.1.3 **Classifications lists for AI model image analysis:** The image detection and semantic models used for image analysis have been trained on a specific range of classifications which limits the tags they can apply. The image detection model is trained on a list of 1,000 classifications using the ImageNet Class List[4]. The semantic model has been trained on a narrower set of 99 classifications available through the COCO (Common Objects in Context) dataset[5]. As noted above participants were not, however, restricted to the same list of classifications and could apply any labels they chose.

4.2 AI prototype results

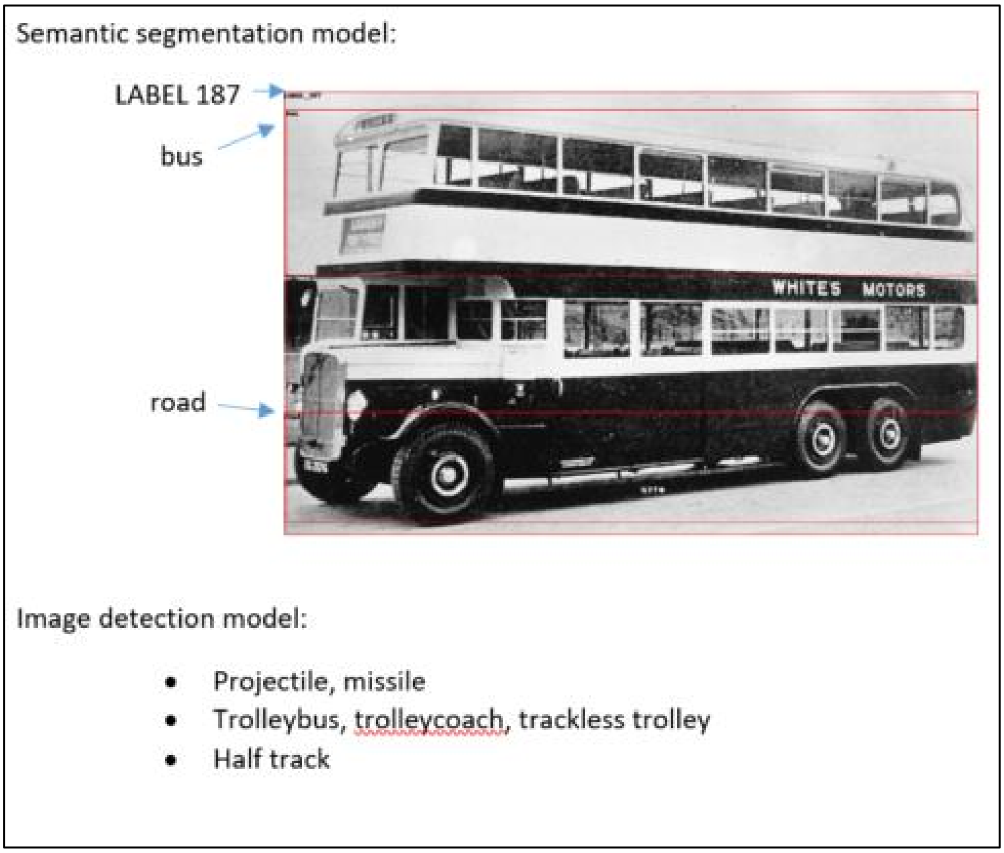
Below are the results of the AI prototype for image analysis of Image 1.[6] The photographic image surrounded by red bounding boxes is the original output from the semantic segmentation model and has been annotated to more clearly indicate the labels for each box. Below the image are the results returned by the image detection model.

[3] Original image source: Barry Library, 'A Whites Motors Six Wheeler Bus'. Available at: https://www.peoplescollection.wales/items/1884606. Licence: Creative Archives Licence.

[4] 'IMAGENET 1000 Class List'. Available at: https://deeplearning.cms.waikato.ac.nz/user-guide/class-maps/IMAGENET/

[5] COCO Dataset. Available at: https://tech.amikelive.com/node-718/what-object-categories-labels-are-in-coco-dataset/

[6] Note: 'LABEL 187' indicates the semantic segmentation model recognised an object in the image but the corresponding tag was not among the classifications list which it was trained on.



Semantic segmentation model:

Image detection model:

- Projectile, missile
- Trolleybus, trolleycoach, trackless trolley
- Half track

Figure 4: Semantic Segmentation and image detection model outputs (Image 1)

4.3 Participant results

4.3.1 **Table 1** on the page below shows the results provided by the AI model and all groups from both the symposium and online testing sessions. Similar instances of the same word or term have been grouped together to show both the range and frequency of the terms used. Labels marked with an asterisk (*) under 'AI Prototype Results' are the results of the semantic segmentation model; those without are the results of the image detection model.

4.3.2 Key figures:

- The total number of terms applied by participants across both sessions was 79
- The average number of terms selected by participants in both sessions was 15.6
- The average for session 1 was 10.6
- The average for session 2 was 23
- The number of unique terms selected by participants (i.e. by one group) was 33
- The number of non-unique terms selected (i.e. by two or more groups) was 15
- Excluding the repetition of closely conceptually related but differently worded labels (e.g. "black and white" and "black and white photograph") the most frequently selected terms by participants were:

− Whites Motors	(5)
− Cardiff	(4)
− Bus	(4)
− Double-decker	(4)

Windows	(4)
Public transport	(3)
Wheels	(3)

- Among the most frequently selected terms by participants, the semantic model applied the term ‘bus’
- The semantic segmentation model also applied the term ‘road’ which was selected by one group in the online testing event
- None of the terms selected by the image detection model were included among participants’ responses
- The term ‘White’s motors’, which was included in all participant group responses, was not picked up by the semantic or image detection mode, as it does not include text analysis, but was an output when the prototype’s text detection function was run over the same image. The text was detected as ‘whitets mottors’ which, although not spelled correctly, is identifiable as the text on the front of the bus.

AI prototype results	**Participant responses**				
	Session 1: AI Research Hub (in person)			***Session 2: Online session***	
	Response 1	**Response 2**	**Response 3**	**Response 4**	**Response 5**
			Black & white	black and white	
		Black and white photograph			black and white photograph
Bus*		Bus	Bus	bus	bus
	Cardiff	Cardiff	Cardiff	cardiff	Cardiff
	Classic				
				coach	
		Destination board			
					door handle
					double axle
		Double-decker	Double-decker	double-decker	double decker
	Double-decker bus	Double-decker bus			
	Empty				
		Empty bus			
					flat tyre
	Four wheels				
					grille
Half track					
					head lamp
	Head lights			headlights	
				industry	
					mud flap
				number plate	number plate
	Old				
				old photo	
		Photograph			
				pontypridd	Pontypridd
Projectile, missile					
			Public-transport	public transport	public transport
Road*					road
					shadow
					seats
			Six-wheeler		
					South Wales transport
				transport	
				transit	
				travel	
Trolleybus, trolleycoach, trackless trolley					
	Twin axle				
	Tyres		Tyres	tyres	
			Vintage		
		Vintage bus			
				vehicle	
				wales	
	Whites motors	White's motors	Whites motors	whites motors	White's Motors
	Windows		Window	windows	windows
					windscreen wipers
					wing mirror
					wheel arches
			Wheels	wheels	wheels
				0770	0778
					1950's

Table 1: AI Prototype and Participant Responses for Image 1 Classification

4.4 Participant grading of AI results

4.4.1 Table 2 below displays the grades that were awarded to the image detection and semantic models for Image 1.

	Image analysis accuracy ratings					
	Session 1 (AI Symposium)			Session 2 (online)		
	Accuracy rating	Number of respondents	% respondents	Accuracy rating	Number of respondents	% respondents
Image detection model	10%	3	100%	0-19%	5	38.5%
	20%	0	0%			
	30%	0	0%	20-39%	5	38.5%
	40%	0	0%			
	50%	0	0%	40-59%	3	23%
	60%	0	0%			
	70%	0	0%	60-79%	0	0%
	80%	0	0%			
	90%	0	0%	80-100%	0	0%
	100%	0	0%			
Semantic segmentation	10%	0	0%	0-19%	0	0%
	20%	0	0%			
	30%	0	0%	20-39%	1	8%
	40%	0	0%			
	50%	0	0%	40-59%	1	8%
	60%	1	33.0%			
	70%	0	0%	60-79%	4	33.3%
	80%	1	33.3%			
	90%	0	0%	80-100%	6	50%
	100%	1	33.0%			

Table 2: Accuracy ratings for image analysis (Image 1)

Average accuracy ratings:

Image detection model	Session 1 (symposium):	10%
	Session 2 (online):	26.92%
Semantic segmentation model	Session 1 (symposium):	80%
	Session 2 (online):	75%

4.5 Analysis and feedback

4.5.1 Results analysis:

Image labelling: There was a difference between the average number of labels applied between participants in each testing session. Symposium attendees in the first session provided an average of 10, while participants in the online session provided an average of 23. The granularity and the technical nature of the responses returned in session 2 was also notable with terms such as ‘door handle, ‘wheel arches’, ‘double axle’ and ‘mud flap’ being used.

Accuracy rating: Participants in session 1 (the symposium) awarded the image detection model an average of 10% and those in session 2 gave it an average of 26.92%. This contrasts sharply with the results for the semantic model, which was awarded an average of 80% among participant groups in session 1 and 75% among participants in session 2. This lower grades awarded to the image detection model are understandable given that none of the object labels it selected were reflected in any of the participants’ responses. The object label ‘projectile/missile’ returned by the image detection model was also noted as a surprising result by participants in both sessions and this may have influenced their perceptions of its accuracy.

4.5.2 Participant feedback:

Session 1: When assessing the accuracy of the AI prototype’s results for semantic segmentation and image detection, some symposium attendees queried their ability to provide a rating without being able to see the classifications lists the model had been trained on. There was also some discussion over whether each model should be judged only against the accuracy of the results returned, or whether the absence of classifications for objects that were clearly visible within the image should be considered too. In the event, it was suggested that participants could factor in the absence of anticipated classifications if they felt this would influence their perception of its accuracy, but it is possible that in future testing participants might benefit from specific guidance on whether to assess the absence as well as presence of classifications.

Session 2: In session 2 participants’ comments included their surprise at some of the image detection model’s outputs, including ‘missile/projectile’ in particular, as well as querying which some of the larger text in the image, such as ‘White’s Motors’, had not been picked up by either the image detection or semantic models. As noted above, a slightly misspelled version of this term was in fact picked up by the text detection (text detail) model, but this is not a feature of the image analysis function.

5. Activity 2: Image segmentation and labelling (Image 2)

5.1 Activity design

5.1.1 **Session 1**: Each group was provided with a printed copy of a black and white photograph of two horses ploughing a field (Image 2). They were requested to draw boxes around objects within the image and to label the contents of each box, emulating the semantic model functions of the AI prototype.[7] In discussion in the round following the activity, participants were also requested to provide an indication of the accuracy level they would award to the image detection and semantic models for their outputs. For the purposes of this discussion on accuracy, the participants were given a printed list of the classifications in the COCO dataset used to train the semantic model, but were not provided with the ImageNet classifications for the image detection model as, at 1000 words, this was considered too long to be usable within the activity timeframe.

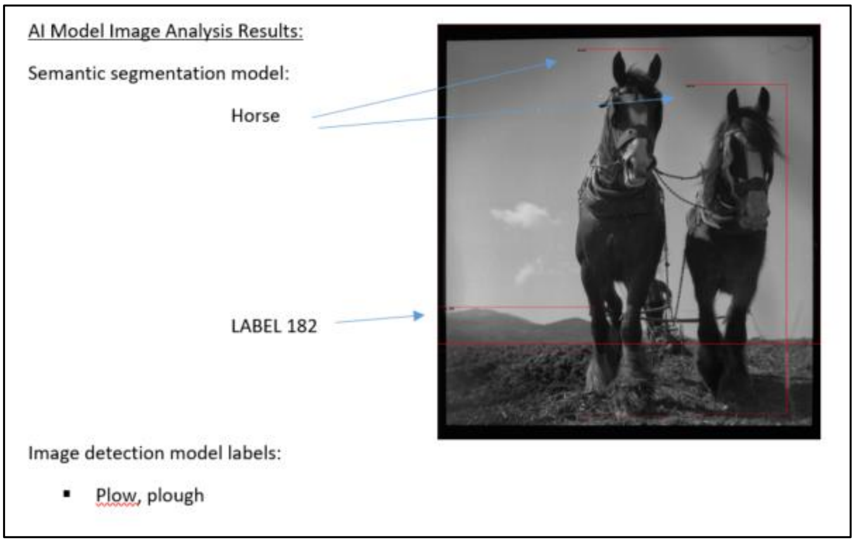
5.1.2 **Session 2**: Each breakout group was requested to complete the activity in the Zooniverse activity webpage. This included a copy of the original image on which participants could draw bounding boxes and label their content. Due to the time constraints of the session, participants in session 2 were not requested to provide an accuracy rating for the AI model's results following the activity, but did provide general feedback in an in-the-round discussion (see section 5.4.2).

5.1.3 **Classification lists**: The semantic segmentation model and image detection model were trained on the same classification lists described in section 4.1.3 above. As in the first activity, participants were not provided with the classifications list when carrying out the segmentation and labelling tasks.

5.2 AI prototype results

5.2.1 Figure 5 below shows the results of the AI prototype for the semantic segmentation and image analysis models for Image 2. The photograph surrounded by red bounding boxes is the original output from the semantic segmentation model, which has been annotated to more clearly indicate the labels for each box. Below this are the results returned by the image detection model.

[7] Original image source: Harris Walter, 'Ploughing with horses in the Conwy Valley'. Available at: https://www.peoplescollection.wales/items/1883486. Licence. Creative Archives Licence.



AI Model Image Analysis Results:

Semantic segmentation model:

Horse

LABEL 182

Image detection model labels:

- Plow, plough

Figure 5: Semantic Segmentation and image detection model outputs (Image 2)[8]

5.3 Participant results

5.3.1 Figure 6 on the page below shows the responses returned by participants in testing sessions 1 and 2.

5.3.2 Key figures include:

- The **total number of terms** applied by participants across both sessions was **52**
- The **average number of terms** selected by participants in both sessions was **10.4**
- The average for session 1 was 6.3
- The average for session 2 was 16.5
- The **number of unique terms** selected by participants across both sessions was **10**, although some of these terms related to the same object within an image which was differently labelled (e.g. mountain/hill; blinkers/blinders)
- The **number of non-unique terms** selected by participants across both sessions was **41**
- The most frequently selected terms by participants were:

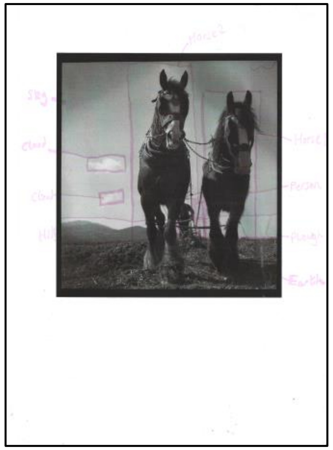
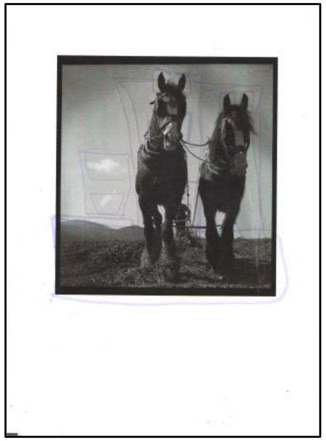
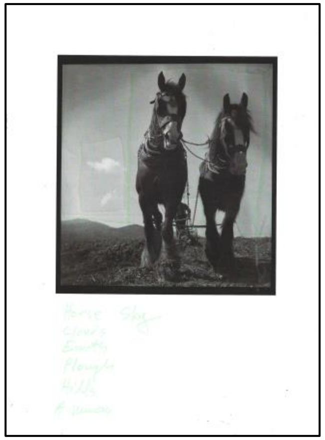
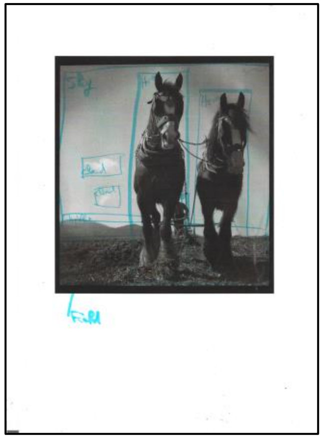
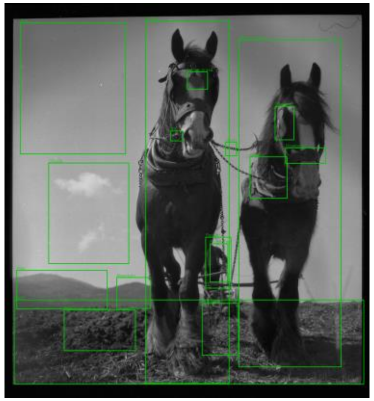
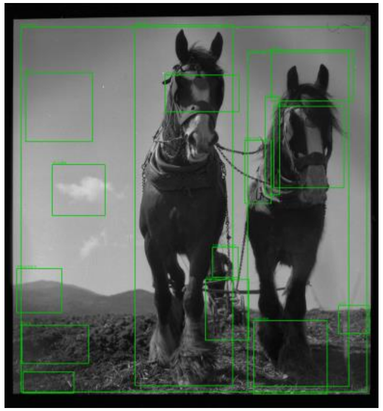
Sky	(5)
Cloud(s)	(5)
Horse	(5)
Hill(s)	(4)
Earth	(3)
Plough	(3)

- Among the most frequently selected terms, the semantic segmentation model applied the term 'horse' and the image detection model selected the term 'plough/plow'.
- Among the segmentation boxes applied by the semantic model, boxes surround the objects that appear to be hills or mountains (LABEL 182) and the ground/earth (label not visible); it is therefore possible that the semantic model actually identified 3 of the objects most commonly identified by our participants, but only linked these to a dataset label in one case.

[8] 'LABEL 187' indicates the semantic segmentation model recognised an object in the image but the corresponding tag was not among the classifications list it was trained on.

- One group in session 1 provided segmentation boxes around the image but did not provide labels for their contents

Figure 6: Participants' responses to image labelling and segmentation task (Image 2)

Session 1	 Sky, cloud, hill, earth, horse, Person	 [None]	 Horse, horse, clouds, earth, plough, hills, [illegible], sky	 Sky, field, cloud, hills, horse
Session 2		Sky, Horse, Horse brass, Shire Horse, Blinders, Noseband, Collar, Person, Ploughing, Plough, Mountain, Hills, Field, Soil, Clouds, Bit		Farming, horse, sky, blinkers, shire horse, mane, ears, bridle, harness, clouds, ploughing, plough, hooves, mountain, field, soil, earth

	Session 1 (AI Symposium)		
	Accuracy rating	**Number of respondents**	**% respondents**
Image detection model	10%	0	0%
	20%	0	0%
	30%	0	0%
	40%	0	0%
	50%	0	0%
	60%	0	0%
	70%	3	100%
	80%	0	0%
	90%	0	0%
	100%	0	0%
Semantic segmentation model	10%	0	0%
	20%	0	0%
	30%	0	0%
	40%	0	0%
	50%	1	33%
	60%	1	33%
	70%	0	0%
	80%	0	0%
	90%	0	0%
	100%	1	33%

Table 3: Accuracy ratings for image analysis (Image 2)

Average accuracy ratings:

Image detection model	Session 1 (symposium):	70%
Semantic segmentation model	Session 1 (symposium):	70%

5.4 Analysis and feedback

5.4.1 Results analysis

As was the case with Image 1, there was again a difference in the level of granularity and technical detail supplied in the responses by participants in session 1 and session 2. Participants in session 2 provided a larger number of object levels and a higher level of detail. While there were too many variables between the setup of sessions 1 and 2 to determine why this may have been the case, relevant factors could include:

- Differences in the professional experiences and knowledge bases of the participants in each session
- Differences in the setup of the activity with the in-person participants using pen and paper to provide their answers and online attendees using the Zooniverse platform
- Differences in the dynamics of online breakout room discussions vs. in-person group discussions

5.4.2 In future sessions, when using human-generated responses to determine the comprehensiveness of the AI model's results, it may be useful to provide more specific instructions on classifications tasks, such as guiding participants to only classify the main objects within image (e.g. a bus) and exclude the components of the object (e.g. window), if this is not how the AI model is expected to operate itself. This may depend on the level of granularity aimed for in the AI model's results. If AI-generated metadata is expected, or allowed, to be more granular than that of a typical archival catalogue or research dataset record, then encouraging more detail may be preferable. Providing guidance on the basis of this decision would help to ensure that human-generated feedback which may be used as a benchmark to assess the AI model's accuracy is consistent among testing groups and aligned with the AI's functions.

5.4.2 **Participant feedback:**

Session 1: When reviewing the outputs of the AI prototype, one participant in session 1 expressed surprise that the semantic model generated the term 'horse' rather than 'plough/plowing', which the image detection model had generated. The participant noted they would have expected this to be the other way around as ploughing was regarded as a better overall description of what was happening in the image, rather than a horse which is only an object contained within it. Another participant suggested, however, that the plough was barely visible in comparison with the two horses pulling it, so it was understandable that the semantic model selected 'horse' instead.[9]

Session 2: Participants expressed interest in why each model had picked up only one object within the image to classify, although in fact the semantic model had identified at least three objects but had not labelled them all. In a reversal of the perspective given in session 1, there was discussion as to why the image detection model picked up the term 'plough' as it is a relatively small object within the image, though another participant suggested that the object of the plough machinery itself could be taken together with the horse to depict one whole object.

[9] The classifications list for the semantic model (COCO dataset) is currently limited to 99 terms, and 'plough' was not included among them. The image detection list (ImageNet) has a larger corpus of 1,000 terms which does include 'plow/plough'.

6. Activity 3: Text summarisation (Text 1)

6.1 Activity design

6.1.1 **Session 1:** Symposium participant groups were provided with a printed photographic image of a programme of events for an opening ceremony event (Text 1).[10] They were requested to write a summary of the programme together and were given no other specific instructions.

6.1.2 **Session 2:** Participants in the online event were requested to work in groups in breakout rooms to complete the activity online via Zooniverse. The online activity provided a copy of the image and requested to write a summary of the document in a free-text box.

6.2 AI prototype results

The AI text summarisation model provided the following summary of the text:

“11.145 a.m. Opening Ceremony111.120a. m. Arrival of special train from London. 4.20departuretrain.ofspecialforp.london1.10 p.M.Civic Reception and Luncheon at the City Hall”

6.3 Participant results

Below are the verbatim answers provided by of each group of participants for the text summarisation task in sessions 1 and 2:

Session 1:

- Answer 1:
  - “Event programme for dedication held between 1902 and 1952. Event includes Rt. Hon so-and-so. Event features travel to and from London and lunch. Event is civic in nature. Event takes place between 11.20am and 4.20pm”
- Answer 2:
  - “Old programme of a posh ceremony”
- Answer 3:
  - “Programme for an opening ceremony, dedication and civic reception.”
- Answer 4:
  - “Programme for an event from 11:20 – 4:20 with a service of dedication and a civic reception”

Session 2

- Answer 1:
  - “A programme of events about a dedication ceremony by the Viscount Cecil of Chelwood”
- Answer 2:

[10] Original source: Welsh League of Nations Union/UNA Wales, ‘Programme of events for Temple of peace opening ceremony ‘. Available at: https://www.peoplescollection.wales/items/1887331. Licence: Creative Archive Licence.

- "Programme for a service of dedication delivered by The Rt. Hon. Viscount Cecil of Chelwood, K.C followed by a Civic Reception and Luncheon at the City Hall. Transport provided by a special train from and to London arriving at 11:20am and departing at 4:20pm."

6.4 Analysis and feedback

6.4.1 **Results analysis:** Notably absent from the AI model's summary was the word 'programme' as a description of the document itself. The word 'programme' is featured at the top of the text itself and featured in every group response. In discussing this, some groups expressed surprise that the word was not featured, though one participant in the symposium testing session noted that the summary model may be trained to summarise the content of the text itself, rather than the title, and it would not therefore be expected to include this word in a summary.

6.4.2 The groups in both sessions were provided with a very broadly defined task of summarising the text, and were not provided with any guidance as to the words they could use to do this. Whereas the AI prototype's summary is restricted to the using the words which it identifies within the text itself, participants' responses included a broader range of words - examples include 'event', 'old', 'posh' and nature'. When asked about this, participants confirmed that they would not have restricted themselves to writing a summary based on the words used in the text itself unless they were explicitly asked to do so.

6.4.3 **Participant feedback on the activity:** Participants in both sessions discussed whether or not they were allowed to look up additional information online to add to the summary of the document and provide context to this. One group in session 1 noted that they had used the internet and had identified the image online on the *People's Collection Wales* website. Participants in session 2 noted that they would have looked up additional information if they thought they were allowed.

6.4.4 **Feedback on the AI model results**: In session 2, one participant asked why the AI model had ordered the text in the summary as it had, noting that it appeared to have prioritised some elements of the text (e.g., "11.45") when the text itself does not appear to suggest any priority for this term. In response a project software engineer clarified that the model might struggle to provide a summary of a short piece of text (which we selected for use in the testing sessions for brevity of reading) as the way that the summary model interprets the structure of text is based on the longer texts, including large paragraphs of text, on which it has been trained.

7. Activity 4: Importance Labelling (Text 1)

7.1 Activity design

7.1.1 **Session 1:** Symposium participant groups were provided with a printed copy of the same programme of events for an opening ceremony event used in Activity 3 (see section 6.1) and a list of the classifications used for ontological importance labelling by our AI prototype.[11] To apply tags to the text, participants used the set of ontological classifications that the AI model uses - the 18-class Named Entity Recognition (NER) model developed by *Flair.* These classifications are listed in Table 3 below.[12] Participants were requested to draw boxes around the words labelled and link these one of the classifications from the list provided.

7.1.2 **Session 2:** Participants in the online event were requested to work in groups in breakout rooms to complete the activity via Zooniverse. The webpage provided a copy of the text and participants were able to draw boxes to highlight parts of the text. Each box drawn produced a drop-down list of the NER classifications to select from.[13]

Tag	Meaning
CARDINAL	Cardinal value
DATE	Date value
EVENT	Event name
FAC	Building name
GPE	Geo-political entity
LANGUAGE	Language name
LAW	Law name
LOC	Location name
MONEY	Money name
NORP	Affiliation
ORDINAL	Ordinal value
ORG	Organization name
PERCENT	Percent value
PERSON	Person name
PRODUCT	Product name
QUANTITY	Quantity value
TIME	Time value

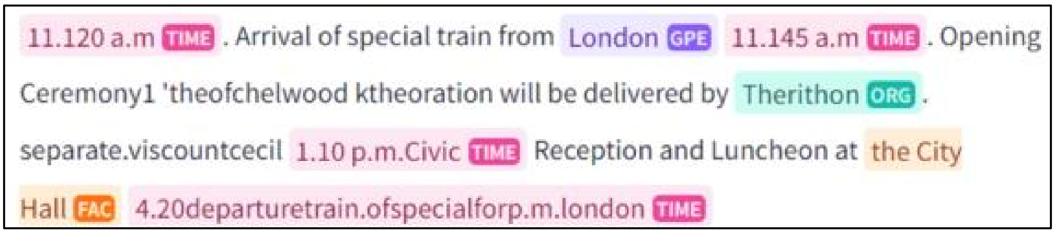
Table 3: Flair Named Entity Recognition classifications (18 class model)

7.2 AI prototype results

7.2.1 Figure 7 below shows the output of the AI prototype's ontological importance classification function in its original output format. The words selected by the model are highlighted together with the relevant Flair NER classification. Figure 8 includes a diagram showing these classifications as applied to the original text source for comparison against participant-generated responses.

[11] 'English NER in Flair (Ontonotes large model)': https://huggingface.co/flair/ner-english-ontonotes-large?text=On+September+1st+George+won+1+dollar+while+watching+a+concert.

[13] Note: In session 1, most groups used the 'tag' (e.g. GPE) to label the words they had highlighted, whereas in session 2 the Zooniverse environment required the corresponding 'meaning' label associated with each tag to be selected (e.g. Geo-political entity). For the purposes of comparison, both sets of results shown in section 7.2 below refer the 'meaning' label. Appendix A include the original responses of groups from both sessions which show the tag or meaning value as originally written.



11.120 a.m TIME . Arrival of special train from London GPE 11.145 a.m TIME . Opening Ceremony1 'theofchelwood ktheoration will be delivered by Therithon ORG . separate.viscountcecil 1.10 p.m.Civic TIME Reception and Luncheon at the City Hall FAC 4.20departuretrain.ofspecialforp.m.london TIME

Figure 7: AI Model Output for Text Importance Labelling (Text 1)

7.2.2 Among the labels selected by the AI model:

- 4 time value (TIME) classifications were applied
- 1 geo-political entity (GPE) was identified
- 1 organisation was (ORG) identified
- 1 building name (FAC) was identified

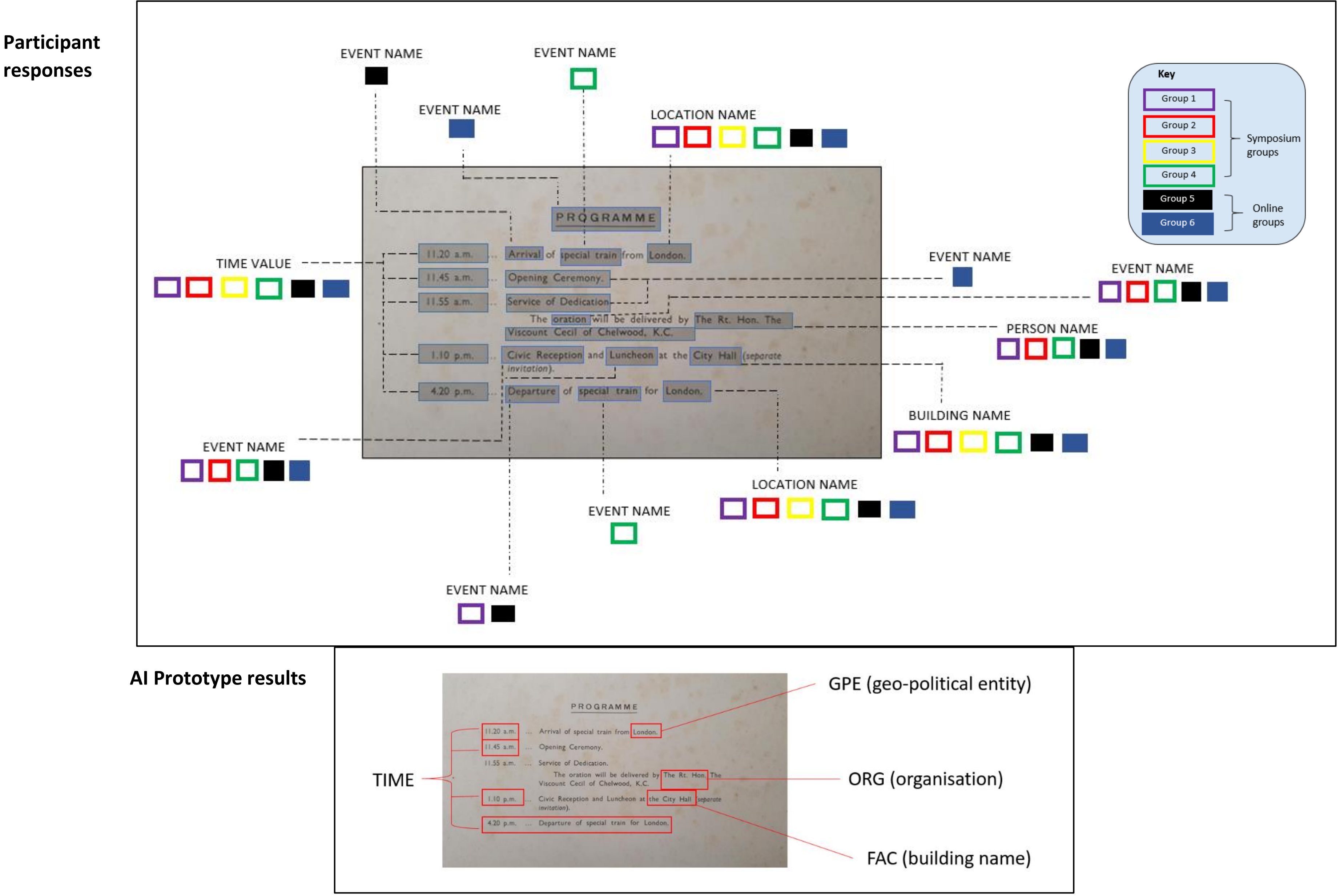
7.3 Participant results

7.3.1 Figure 8 below shows the Flair NER categorisations that were applied by participant groups in testing sessions 1 and 2 alongside a diagram of the results provided by the AI ontological classification model. The number of squares next to each label for the participant results indicates how many groups applied the label to each term highlighted.

7.3.2 Among the results shown, key figures include:

- The total number of classifications applied by the AI model was 7
- The **average number** of classifications applied by **participant groups** was **11.6**
- The **average number** of classifications applied by groups in **session 1** ("symposium groups" below) was **13**
- The **average number** of classifications applied by groups in **session 2** ("online groups" below) was **11**
- The **total number of classifications** applied by one or more **participant groups** across both sessions was **17**, including:
  - 5 time values
  - 8 event names
  - 2 location names
  - 1 person name
  - 1 building name

Figure 8: AI Prototype and Participant Responses: Importance Labelling (Activity 4)



Participant responses

AI Prototype results

PROGRAMME

11.20 a.m. ... Arrival of special train from London.

11.45 a.m. ... Opening Ceremony.

11.55 a.m. ... Service of Dedication.

The oration will be delivered by The Rt. Hon. The Viscount Cecil of Chelwood, K.C.

1.10 p.m. ... Civic Reception and Luncheon at the City Hall (*separate invitation*).

4.20 p.m. ... Departure of special train for London.

7.4 Analysis and feedback

7.4.1 **Results analysis:**

- Event name label: There was a notable difference between the AI ontological classification results and participants' responses for the classification 'event name' in particular, as it was not applied at any words at all by the AI model, but was applied to eight terms in participant group's responses. The participants also appeared to differ in their views about what constitutes an event as in five of these instances the event name label was only applied by one group.
- Location vs Geopolitical Entity classification: London was recognised by the AI model as a geo-political entity, whereas participant groups unanimously labelled the city with 'location name', and this was discussed in participants' discussions following the task (see section 7.4.2 below).
- Organisation label: The AI model notably picked up the words 'Right Honourable' as an organisation and did not associate this with a person's name, in contrast to the participants' responses. It is unclear why the AI model applied the organisation name classification in this event, but it is possible that the title is associated with Parliament as an organisation.
- Correlation between AI and participant results: Overall, the AI model's outputs correlated in 5 instances with participants' responses (four-time values and the building name), and differed in two instances (geo-politial entity and organisation)

7.4.2 **Participant feedback:**

Post-activity discussion and feedback from symposium participants included:

Session 1

- Discussion over the application of the tag geo-political entity classification to London; participants all agreed they would be more likely to apply the location tag as they did not consider London a geo-political entity, but agreed it was understandable for the AI model to select this. Participants suggested that London was not sufficiently large to constitute a geographical area with a political identity or structure distinct from the wider UK.
- One participant also noted that, while the AI model's classifications appeared relatively accurate, fewer were identified than they anticipated.

Session 2:

Feedback on the activity setup included:

- The task was 'easy enough' but the list of ontological classifications was felt to be slightly limited as one group had identified more words for importance labelling if appropriate classifications were available
- The way in which the bounding boxes were setup in the online activity meant that terms which would be classified together as a unit but which run over one line of text were difficult to distinguish from the other words surrounding them

Feedback on the AI model's classifications: One participant suggested that they could understand why the AI ontological function applied fewer classifications because of how the text was firstly interpreted by the AI model (shown in Figure 7 above). This raises the question of whether the text which participants classify should be the original text source, or the text as interpreted by the AI model. While the former might give a clearer idea of how a person would approach the activity from scratch, using the AI-interpreted text might be a better way of determining the accuracy of the ontological classification function as distinguished from text interpretation (text detail model) itself.

8. General feedback

8.8.1 Towards the end of each session, participants were invited to provide their general feedback on their experience of joining in the testing activities. There responses included:

Session 1:

- Participants agreed the length of the session, and number of activities was acceptable, and that they had enough time given to complete the tasks requested.
- It was indicated that a session which lasted more than two hours might be too long.

Session 2:

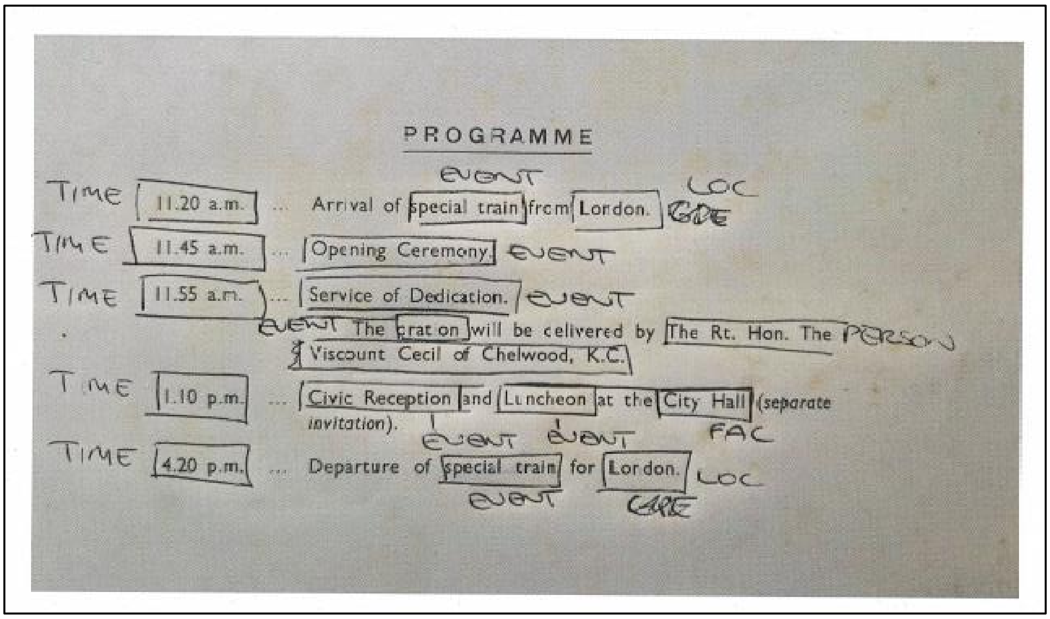
- Participants in session 2 also felt that an appropriate amount of time was given for the tasks and that the breakout rooms used for activities facilitated smaller group discussions about the activities and more interaction than is possible in full group discussions.
- One participant wondered whether we could consider individually based testing activities rather than group-based activities; a project leader responded that the aim of the session was to generate interaction and discussion as much as to assess the model's accuracy.

Appendix: Original Group Responses for Activity 4

(Importance Labelling – Text 1)

Session 1

Response 1: Classifications applied: Time, Event, Location, Person



PROGRAMME

11.20 a.m. ... Arrival of special train from London.

11.45 a.m. ... Opening Ceremony.

11.55 a.m. ... Service of Dedication.

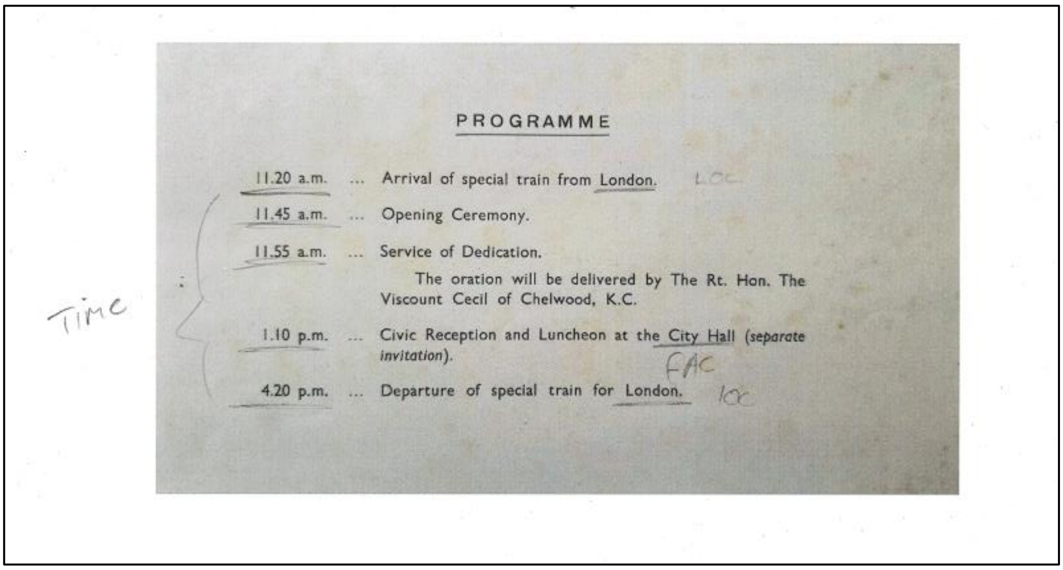
The oration will be delivered by The Rt. Hon. The Viscount Cecil of Chelwood, K.C.

1.10 p.m. ... Civic Reception and Luncheon at the City Hall (*separate invitation*).

4.20 p.m. ... Departure of special train for London.

TIME
TIME
TIME
TIME
TIME
EVENT
LOC
GPE
EVENT
EVENT
EVENT
PERSON
EVENT
EVENT
FAC
LOC
EVENT
GPE

Response 2: Classifications applied: Time, Building Name, Location



PROGRAMME

11.20 a.m. ... Arrival of special train from London.

11.45 a.m. ... Opening Ceremony.

11.55 a.m. ... Service of Dedication.

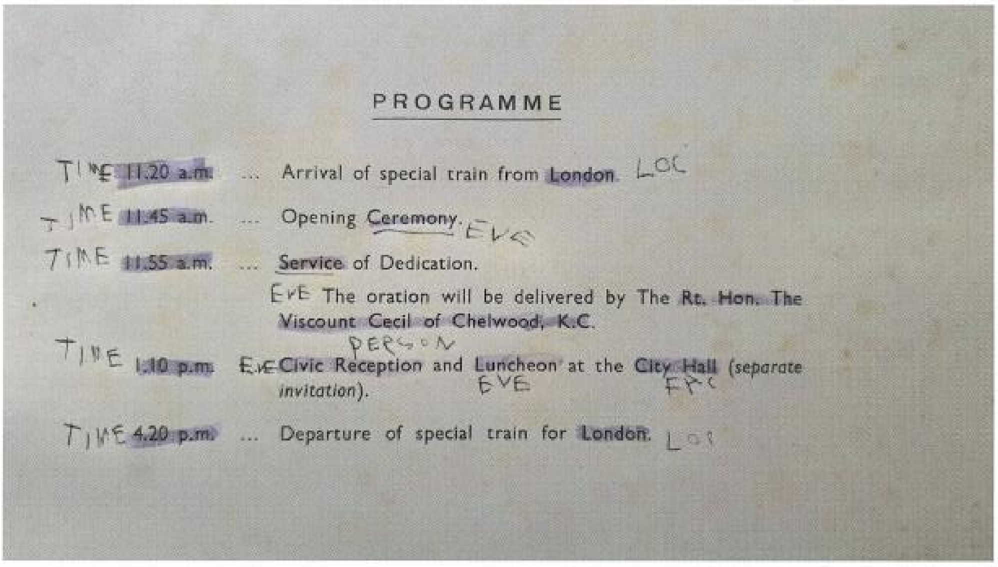
The oration will be delivered by The Rt. Hon. The Viscount Cecil of Chelwood, K.C.

1.10 p.m. ... Civic Reception and Luncheon at the City Hall (*separate invitation*).

4.20 p.m. ... Departure of special train for London.

Time
LOC
FAC
loc

Response 3: Classifications applied: Time, Location, Event, Building Name, Person



PROGRAMME

TIME 11.20 a.m. ... Arrival of special train from London. LOC

TIME 11.45 a.m. ... Opening Ceremony. EVE

TIME 11.55 a.m. ... Service of Dedication.

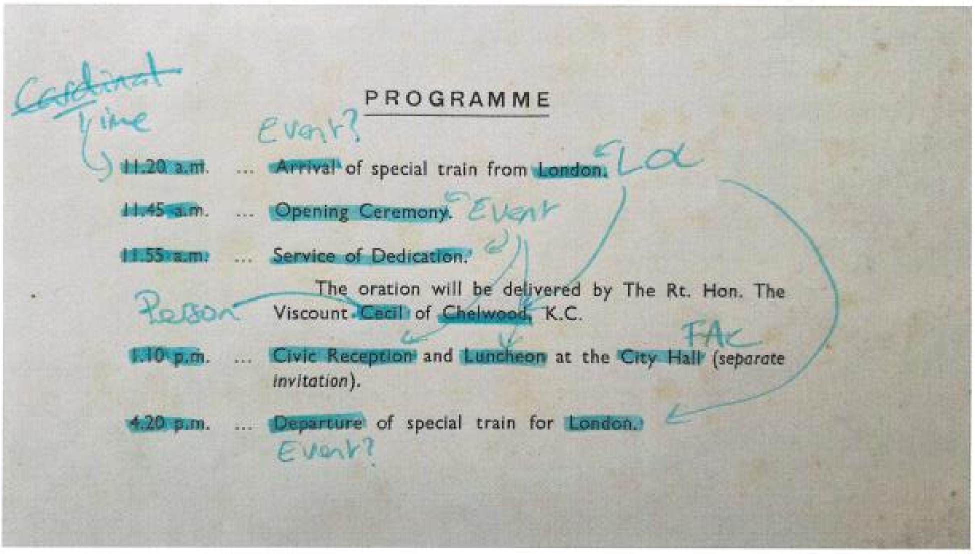
EVE The oration will be delivered by The Rt. Hon. The Viscount Cecil of Chelwood, K.C.

PERSON

TIME 1.10 p.m. EVE Civic Reception and Luncheon at the City Hall (*separate invitation*). EVE FAC

TIME 4.20 p.m. ... Departure of special train for London. LOC

Response 4: Classifications applied: Time, Person, Location, Event, Building Name



~~Cardinal~~ time

PROGRAMME

Event?

11.20 a.m. ... Arrival of special train from London. Loc

11.45 a.m. ... Opening Ceremony. Event

11.55 a.m. ... Service of Dedication.

Person The oration will be delivered by The Rt. Hon. The Viscount Cecil of Chelwood, K.C.

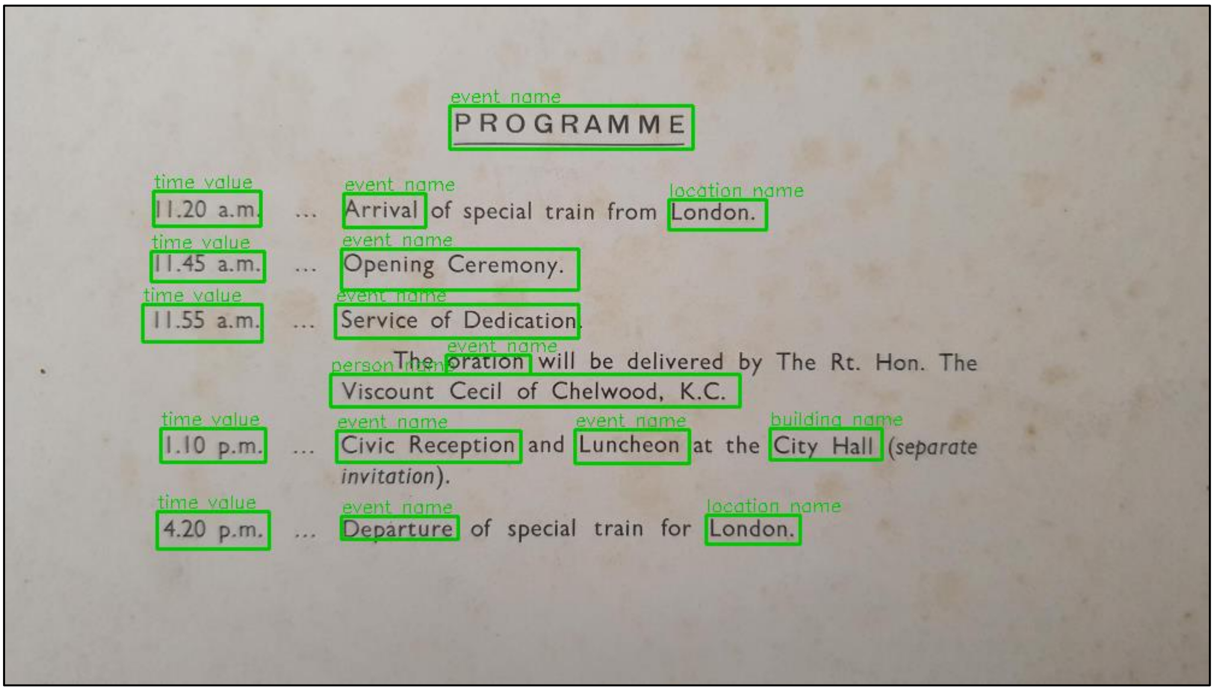
1.10 p.m. ... Civic Reception and Luncheon at the City Hall (*separate invitation*). FAC

4.20 p.m. ... Departure of special train for London.

Event?

Session 2

Response 1: Classifications applied: time value, event name, location name, building name, person name



PROGRAMME

11.20 a.m. ... Arrival of special train from London.

11.45 a.m. ... Opening Ceremony.

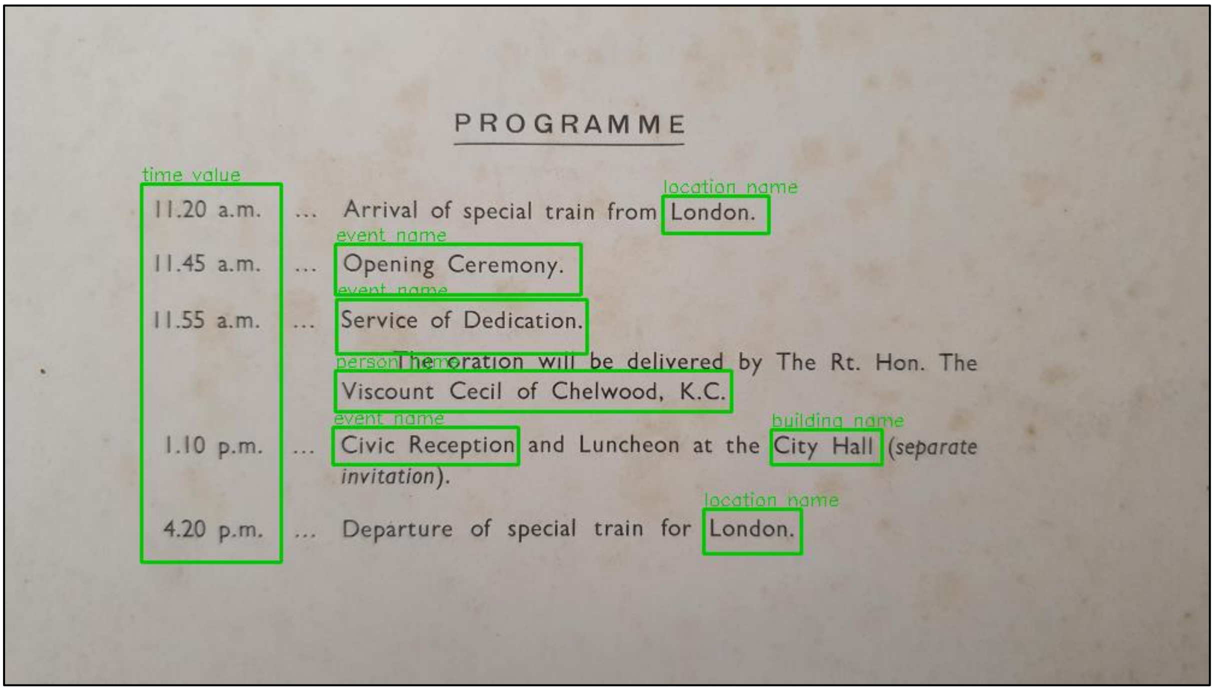
11.55 a.m. ... Service of Dedication.

The oration will be delivered by The Rt. Hon. The Viscount Cecil of Chelwood, K.C.

1.10 p.m. ... Civic Reception and Luncheon at the City Hall (*separate invitation*).

4.20 p.m. ... Departure of special train for London.

Response 2: Classifications applied: time value, location name, building name, event name, person name



PROGRAMME

11.20 a.m. ... Arrival of special train from London.

11.45 a.m. ... Opening Ceremony.

11.55 a.m. ... Service of Dedication.

The oration will be delivered by The Rt. Hon. The Viscount Cecil of Chelwood, K.C.

1.10 p.m. ... Civic Reception and Luncheon at the City Hall (*separate invitation*).

4.20 p.m. ... Departure of special train for London.