

EXPLANE: An Extensible Framework for Poster Annotation with Mobile Devices

Nikhil Gopal¹, Andrew Su², Chunlei Wu², Sean D. Mooney¹

1. Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA
2. The Scripps Research Institute, La Jolla, CA

Abstract

Summary: Scientific posters tend to be brief, unstructured, and generally unsuitable for communication beyond a poster session. This paper describes EXPLANE, a framework for annotating posters using optical text recognition and web services on mobile devices. EXPLANE is demonstrated through an interface to the MyVariant.info variant annotation web services, and provides users a list of biological information linked with genetic variants (as found via extracted RSIDs from annotated posters). This paper delineates the architecture of the application, and includes results of a five-part evaluation we conducted. Researchers and developers can use the existing codebase as a foundation from which to generate their own annotation tabs when analyzing and annotating posters.

Availability: Alpha EXPLANE software is available as an open source application at <https://github.com/ngopal/EXPLANE>

Contact: Sean D. Mooney (sdmooney@uw.edu)

1 Introduction

Although posters are a first-line medium for presentation of novel scientific research, posters constrain authors to a limited amount space. As a consequence of these space constraints, otherwise pertinent information and context are omitted from posters, requiring readers to obtain further information and context from those presenting posters.

Posters in genetics often include genetic variant annotations. Although variants are richly annotated and linked to a variety of bioinformatics resources, these annotations are primarily accessible through a desktop computer. In order to increase accessibility and timeliness of

these annotation information, we provide an extensible poster annotation framework other researchers can build upon--EXPLANE. This paper demonstrates EXPLANE through a mobile application that allows end-users to take a picture of a poster containing RSID variant identifiers, submit the poster for processing, and returns linked information contained in credible resources.

Since smartphones provide a novel touch interface, are portable, and are connected to the Internet, many mobile applications taking advantage of these capabilities have been developed. Mobile applications are generally considered to be timely, contextual, and convenient [1]. Many mobile applications have been built to facilitate tasks that would otherwise require a standalone desktop, such as illustrating macromolecular structure, or analysis results of increasingly complex biological information [2]–[6]. Recently, mobile applications have also been developed to act either directly as an endpoint for query services, or using query services as part of their core functionality, shifting computation from the mobile device to a web server [7], [8]. EXPLANE is a mobile application that queries Optical Character Recognition tools (OCR) and variant annotation services to provide a summary of genetic information.

Much research has been conducted on accurately parsing gene names from scientific texts. Although there are generally accepted conventions with gene names, such as capitalizing all letters for human genes, there are a number of semantic details that complicate this task. For instance, gene name nomenclature is not standardized and genes may have multiple names [9]. In addition, gene names vary in terms of digits, letters, special characters, Greek letters, and some genes have compound names [9]. Other challenges associated with gene mapping are gene mention detection (identifying text referring to gene names), building a gene name dictionary, tokenization (providing a list of variations/synonyms/abbreviations for a gene), string matching (approximate string matching due to variability in naming), and false positive filtering (assessing whether a found gene is actually a disease, ambiguously named, etc.) [10].

A number of methods have been used to successfully extract gene names from text, with improvements and refinements of these methods over time. For example, machine learning approaches such as rule-based systems or support vector machines obtain precision and recall of about 83% and 84% respectively [13], [14]. Since the effect of features seems to be small, machine learning approaches seem to perform the best when using all available features [14]. Approaches such as GeNo and BioTagger, which use a dictionary approach in addition to machine learning, tend to yield slightly better precision and recall results of about 88% and

85%-89% respectively [10], [15]. To improve performance even further, researchers have created approaches targeting certain sub-domains (e.g. expecting genes from yeast only). LeadMine is a more recent approach, in the domain of drug name identification, uses a grammar- and dictionary-driven method has shown promising results with close to 90% precision and 85% recall, although these measures were not calculated using the BioCreative corpus [16]. Another example is pGenN, in the domain of plants, is a dictionary based approach that is able to obtain 91% precision and 87% recall [17].

Although the performance of gene name recognition methods have been steadily improving, the false positive rate remains too high to be able to use in this particular mobile application. Thus, to minimize false positive calls to API services, and to include functionality for all organisms, EXPLANE is designed to parse and use Reference SNP cluster IDs (RSIDs) from poster text.

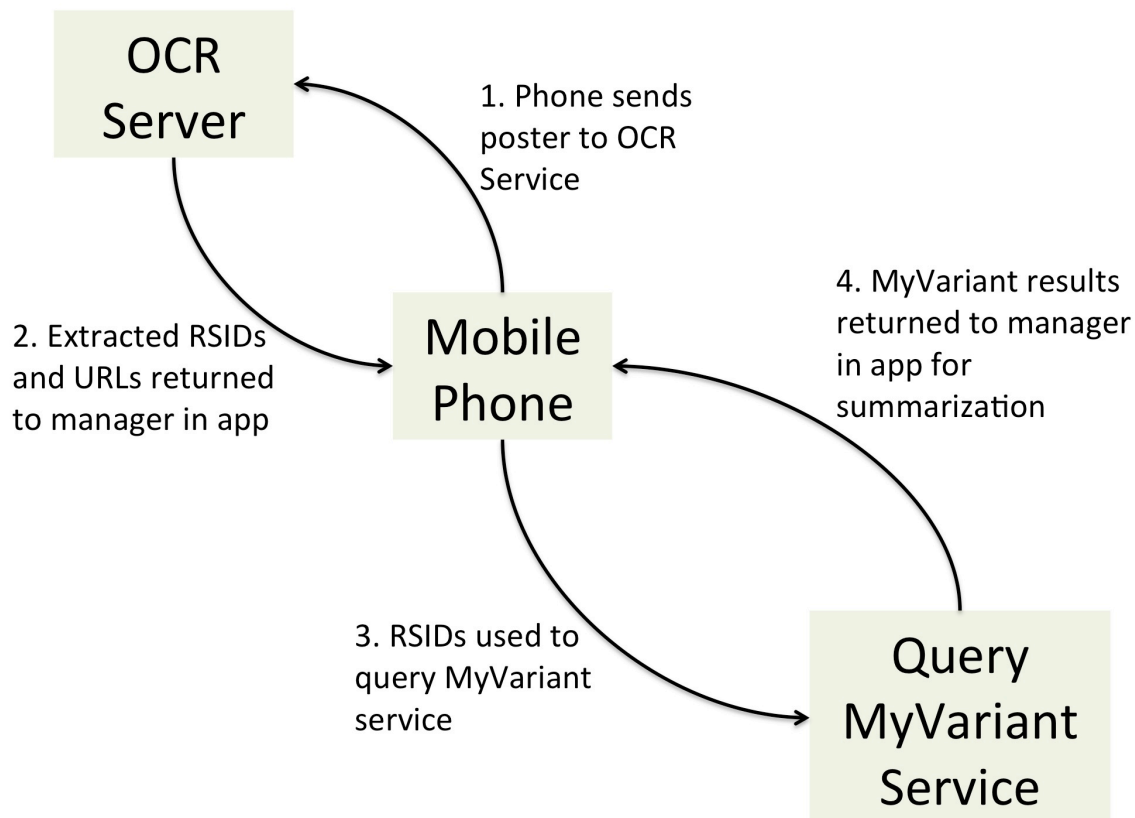
Biological Web Services: Since storage and memory capacities are limited on mobile devices (relative to desktop computers), EXPLANE is designed to use a data federation approach, whereby required data are accessed through web services.

MyVariant.info has an application user interface (API) that is well suited for this task [18]. MyVariant.info serves as an up-to-date repository for variant information, sourcing information from more than 10 well-used bioinformatics repositories [18]. Queries may be made to MyVariant.info using RSIDs, chromosome location, and HGVS nomenclature. Additionally, MyGene.info is another endpoint allowing one to query using gene information, rather than variant information, although MyGene.info specifically is currently not queried in EXPLANE. As of July 2013, MyGene.info endpoint averaged over 3 million requests per month [18]. Furthermore, the MyVariant.info and MyGene.info endpoints are also available through an R package. However, other services, such as BioWAP, also provide access to federated biological resources and structured results, although specifically for protein-related data [19], [20].

2 Workflow and Development

The workflow described for the example application is implemented in Javascript/HTML5/CSS3 and transpiled to iOS and Android using Cordova/PhoneGap software. Specifically, the data binding operations are handled using D3.js [21]. Figure 1 below provides an overview of the workflow.

Figure 1 – An overview of the workflow of EXPLANE. Once a user captures a picture and uploads it to the application, the application sends the poster to an OCR service that parses the text and extracts the RSIDs. Found RSIDs are returned to the application, which creates and submits a query to MyVariant.info. The MyVariant.info service returns a JSON file containing information linked to the RSID, which is summarized and displayed back to the user.



The EXPLANE framework facilitates the following workflow:

- (1) Mobile users provide a photograph of a poster to the application. The photograph should be of reasonably high quality so the Optical Character Recognition (OCR) service can accurately read the contents.
- (2) The application uploads the image to an OCR service, which returns RSIDs. Our OCR server is set to have a generous 10-minute connection timeout, which is more than sufficient to upload, process, and return RSIDs in most cases. It is possible to perform the upload step using mobile data, although we recommend connecting to Wi-Fi.

- (3) The RSIDs are automatically formatted into a URL and are submitted to the MyVariant.info [18] service as a query. EXPLANE is parameterized to only process up to 100 RSIDs, although this parameter may be adjusted to accommodate up to 1000 RSIDs. Currently, the query is executed via a GET request for each RSID, In the future, we will switch to use POST to submit the query for multiple RSIDs in one batch.
- (4) The query results from MyVariant.info are returned to the application in Javascript Object Notation (JSON) format. Before further processing, for the purposes of this particular mobile application, the returned JSON file is inverted (keys in the hash become the values, and the values in the hash become the keys). Since the keys representing specific types of biological information within the returned JSON file are standardized, the JSON may be inverted. For instance, the JSON may be returned with the key 'dbsnp' and the key 'clinvar', and both of those elements may contain a value under 'rsid'. In EXPLANE, this JSON would be inverted such that 'rsid' would be the key, and such that 'dbsnp' and 'clinvar' would be the values. This inversion is what allows EXPLANE to present the same data across an array of biological resources. Without inverting the JSON file, one would have to iterate through a list of resources and check to see if the entry under that resource contains the information one is interested in. By inverting the JSON file, one may call the entry directly, which would return all the relevant sources containing the desired entry.
- (5) The application uses the information contained in the JSON results to present a summary to the mobile users. Although the summary only presents information on RSIDs, gene names, and associated phenotypes, much more information is available for inclusion in the summary.

OCR Service

The OCR service is provided using Tesseract and NodeJS software, and is deployed on a University of Washington production server [22], [23]. Specifically, the server software is written in Express.js, and the Tesseract software is interfaced using the “node-tesseract” javascript library. When a photograph is POSTed to the server, the server uses Tesseract via “node-tesseract.”

Although Tesseract is capable of being trained to read a variety of text, our own Tesseract server is parameterized with default settings, which means that it is optimized to recognize fairly regular blocks of text, rather than single sentences and words, or blocks of text in unexpected

orientations (e.g. in a circle). The following OCR URL accepts post commands:

<https://ocr.iths.org/api/photo>. The source code for a deploy-ready implementation of the OCR service is available at <https://github.com/ngopal/EXPLANE-OCR-Server>

At times, parsed RSIDs may be victim to common character misrecognitions. In order to compensate for these misrecognitions, a post-parsing filter has been implemented to catch and correct common OCR errors. This is possible thanks to the predictable structure of RSIDs (i.e. “rs” followed by a string of between 3 to 9 numeric characters). The common misrecognitions that are corrected for are: “r5” to “rs”, “l” to “1”, “z” to “2”, “o” to “0”, and “e” to “0”.

Genetic Information

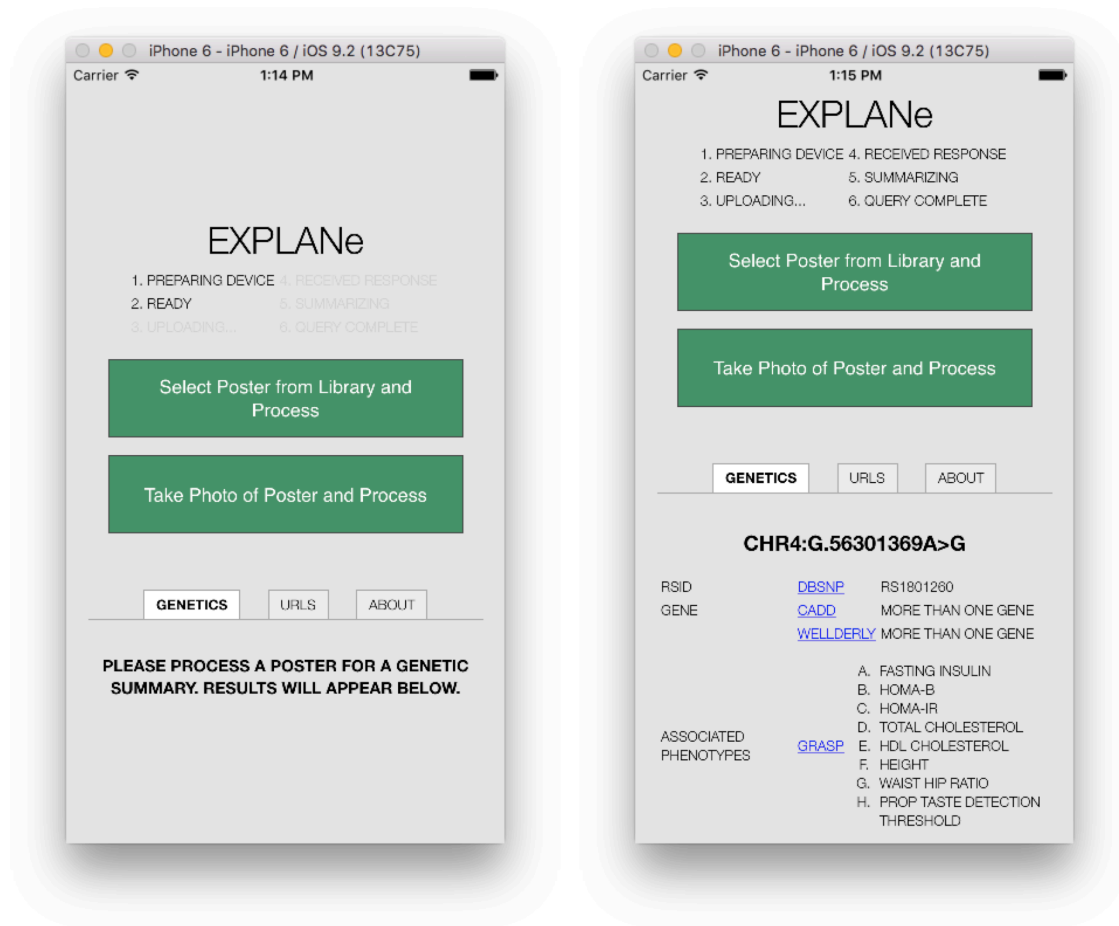
The genetic information presented to the user is obtained from MyVariant.info (<http://myvariant.info>). MyVariant.info is a web service that centralizes genetic information and associated metadata otherwise found across several biological resources. The returned JSON contains much more information than is currently presented to the user in the EXPLANE interface. For instance, in the scenario where gene name is to be displayed, gene name can be found across several resources. Although EXPLANE identifies variants using RSID, each variant that is found in a poster is presented using its HGVS equivalent. Currently, EXPLANE only reports RSID, HGVS ID, gene name, and associated phenotypes (as found in scientific literature).

Application Interface

Figure 2 below presents the interface of the application. In our example, we present three tabs: “Genetics”, “URLs”, and “About”. The Genetics tab is where results from the query process are presented. The “URLs” tab presents the URLs extracted from the poster. The “About” tab is used to provide context and instructions to the user. EXPLANE allows tabs to be easily added or removed.

Figure 2 – An overview of the upload interface and results interface. At the top of the application is a status bar, updating the user about the state of the application. There are six steps in the whole process. If the application is initiated successfully, the first two steps will complete and await user input (i.e. a poster image). Upon uploading, steps three through six may be completed. If the application stops at step four, then something went awry during the parsing process (and the user should be notified via an alert). Below the status bar are two large buttons

the user can use to upload a picture to EXPLANE—either by uploading a pre-existing picture from the photo library on the mobile device, or taking a new poster picture on the spot (which is processed without saving the image to the photo library on the mobile device). Upon successful completion of the process, a summary is displayed under the three tabs.



4 Evaluation

We have evaluated EXPLANE in five different ways to understand the strengths and weaknesses of the application. We have evaluated functionality, effect of font size, effect of RSID lengths, acceptable picture distance from a poster, and consistency of repeated measures. Since the OCR server is setup with default settings, which requires blocks of text of recognition, our test posters are necessarily designed to reflect this property. Please refer to Figure 3 for an example of a test poster.

Figure 3 – The 8.5 inch x 11inch test poster that was used for certain portions of the evaluation. This poster is representative of the type of test poster submitted to EXPLANE during the course of the evaluation.

Variants implicated in b-interleukin cell interaction with riboflavinoids initiating PDP signal cascade

By Someone Sciencey¹ and Another Person²
University of Someplace¹, University of NoOne²

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis venenatis lacus vel egestas semper. Aenean vulputate lorem non velit semper, eu egestas justo porta. Donec aliquam volutpat massa ut sodales. Suspendisse ac risus ipsum. Class aptent taciti rs12210050 ad litora torquent per conubia nostra, per inceptos himenaeos. Nam pulvinar ut eros eget accumsan. Vestibulum id mauris mattis, vehicula felis non, tincidunt elit. Donec porttitor cursus commodo. Fusce neque elit, dictum et consectetur ut, pellentesque quis leo. Rs149183310 quis malesuada purus, id aliquet massa. Phasellus placerat mi risus, vel fermentum metus vehicula a. Etiam laoreet ante ac leo condimentum lobortis. Etiam convallis, augue eget semper pellentesque, neque ex facilisis risus, sed ultrices metus libero ac ligula. Etiam elementum rutrum mi.

Praesent nec leo in sem sagittis posuere. Sed dictum fringilla nibh, eget egestas libero rs55874825 ut. Ut in efficitur diam. Nullam elit nulla, porttitor in hendrerit id, varius vitae arcu. Aenean viverra tristique ultricies. Lorem ipsum dolor sit amet, consectetur rs33980500 elit. Nulla facilisi. Mauris consequat commodo magna in fringilla. Donec finibus volutpat quam, et tempor neque lacinia ut.

Aliquam commodo felis vitae purus imperdiet, in rs3857067 ante eleifend. Nullam dolor eros, egestas tempus venenatis ut, aliquet a nunc. Curabitur venenatis velit id sapien mattis euismod. rs1368011 quis malesuada magna. rs72960926 ornare vehicula dui eu rhoncus. Nulla sit amet arcu lacus. Nam ultrices est a dolor pretium finibus scelerisque non orci. Curabitur et turpis bibendum, molestie risus in, rs77803164 elit. Donec vestibulum blandit leo, non blandit lacus sagittis id. Nulla egestas aliquam leo, eu interdum tortor laoreet quis. Duis vitae molestie est, id gravida tellus. Donec mollis eget arcu in tincidunt. Duis eu diam condimentum, fermentum tortor non, ultrices metus. Suspendisse viverra nisi vitae nunc blandit, et luctus ex condimentum.

felis non, tincidunt elit. Donec porttitor cursus commodo. Fusce neque elit, dictum et consectetur ut, pellentesque quis leo. Curabitur quis rs1057910 purus, id aliquet massa. Phasellus placerat mi risus, vel fermentum metus vehicula a. Etiam laoreet ante ac leo condimentum lobortis. Etiam convallis, augue eget semper pellentesque, neque ex facilisis risus, sed rs2271749 metus libero ac ligula. Etiam elementum rutrum mi.

Variants	Disease	Quant
rs12457997	Diabetes	-0.3
rs6931514	Cancer	4.5
rs166040	Cancer	1.3
rs4778720	AD	7.4
rs1808470	Cancer	10.7
rs401681	Cancer	3.5
rs7904985	HBSIL	7.1
rs4680534	Breast Cancer	6.6

Evaluating Functionality: The functionality of EXPLANE was evaluated in two ways. In order to test the number of posters EXPLANE returns results for, images of various posters containing RSIDs from conferences and web resources were submitted to the mobile application. Of the 11 poster images that were submitted to EXPLANE, 3 of them (27%) returned results. Analysis of the successfully processed conference posters suggests that best results are obtained when capture angle is minimized. Of the 3 successfully processed posters, 1 of the posters contained RSIDs that were misrecognized. The ideal conditions for a poster would be minimized angles (i.e. taking a picture in as centered an orientation as possible), maximizing text size (i.e. being as close to the poster as possible), and ensuring there is sufficient lighting (for exposure), and minimizing blurs in photographs. These results are available in Table 1.

Table 1 – Analysis results for conference posters submitted to EXPLANE. In the “Success” column, “Yes” means all RSIDs were recognized, and “No” means not all RSIDs were recognized. *Two pictures of poster 6 were available, one more distant than the other. **Poster 10 is a negative control as it did not contain any RSIDs at all.

Poster	Success (Yes/No)	Angle
Poster 1	Yes	Subtle Right
Poster 2	No	Subtle Top
Poster 3	Yes	Subtle Left
Poster 4	No	Subtly Left and Rotated
Poster 5	No	Hard Top Right
Poster 6	Yes	Centered
Poster 6 *	Yes	Centered, but distant
Poster 7	Yes	Centered, but distant
Poster 8	No	Subtle Top
Poster 9	No	Centered
Poster 10 **	No	Subtle Bottom

Evaluating Effect of Font Size: In order to evaluate the effect of input text font size on returned results, we created an artificial poster with text in various font sizes and submitted them to EXPLANE. The denominator in the “Number of Correctly Recognized RSIDs” column denotes the number of RSIDs in the poster, rather than the unique number of RSIDs. Pictures of these artificial posters were created in Microsoft PowerPoint, saved as large, high-resolution images, and submitted directly to EXPLANE using an iPhone simulator. Of the 3 submitted poster variations, all 3 returned results. The artificial poster variations were tested at three font sizes—results are available in Table 2. We did not test below a font size of 12, as anything smaller than that size would be too small to reasonably read.

Table 2 – Processing results for posters of varying font size (as submitted to EXPLANE via the iPhone simulator).

Font Size	Number of Correctly Recognized RSIDs
24	4 / 4 (100%)
18	5 / 5 (100%)
12	8 / 8 (100%)

Although RSIDs were recognized correctly 100% of the time under all tested font sizes, URLs were incorrectly recognized, and thus unable to be successfully processed and presented in the application interface. Recognizing URLs is difficult when it is conveyed in small text, as is commonly the case with citations. More challenging still is developing a regular expression pattern that encapsulates all of the required use cases and simultaneously accounts for possible modes of processing error. At the moment, EXPLANE has not implemented the complex logic required to handle URLs that are not easily recognizable and presentable, although the application can be extended to feature this capability.

Evaluating RSID lengths: The purpose of this experiment was to determine if the length of an RSID has an effect on whether it is detected in EXPLANE. A list of human RSIDs were downloaded from UCSC Table Browser (SNP146) and organized by character length of the RSID name [24]. The character lengths of RSIDs ranged from 5-11 (including the “rs” prefix). Due to the skewed distribution of RSID name lengths (please see Table 3), we randomly selected ten RSIDs of lengths 5, 7, 9, and 11, included them in a poster, and submitted the poster to EXPLANE. Randomly selected variants were verified to return query results from MyVariant.info. Poster pictures were taken from a distance of 2 feet. The test poster in this experiment was primarily an enlarged table with nine rows of RSIDs, as well as a title that contained an RSID.

Table 3 – An overview of the distribution of RSID name lengths

Lengths	5	6	7	8	9	10	11
Frequency	59	872	810	66,686	317,790	1,014,092	9,898,529

Table 4 – RSID length versus rate of summarization (found RSIDs / included RSIDs)

RSID Length	5	7	9	11
Summarization Rate	10/10	10/10	10/10	10/10

Based on these results, we determine that variation in RSID name length minimally affects summarization performance.

Evaluating Distance: In this experiment, we captured pictures of posters from various distances, ranging from 1 to 10 feet away from the poster. The first poster used for this experiment was one presented at a Gordon Research Conference (GRC). The GRC poster is 48 inches x 48 inches and contains a table with RSIDs. Due to Gordon Conference rules and restrictions, we cannot provide an image of this particular test poster. The second poster was the 8.5x11 test poster in Figure 3, containing 18 RSIDs spread across text and tables.

Table 5 – Results for Gordon Research Conference poster

Distance	1ft	2ft	3ft	4ft	5ft	6ft	7ft	8ft	9ft	10ft
Recognition	None	Found	Found	Found	Found	Found	Found	Err	Err	None

Table 6 – Results for 8.5 x 11 inch poster

Distance	1ft	2ft	3ft
Recognition	18 / 18	None	None

In Table 5, it may seem peculiar that a picture taken from a distance of 1 foot yielded no results. However, since the table in the GRC poster is 4 feet x 4 feet, and since our Tesseract implementation is parameterized to recognize blocks of text, we think there may not be enough surrounding text for our OCR server to recognize RSIDs. The physical size of the RSID text in the table of the GRC poster is 0.5 inches, and the physical size of the RSID text in the table of

the 8.5x11 poster is roughly 0.2 inches. If we compare the ratios of physical text size to breaking point distance, we see that GRC poster has a ratio of 0.006 and that the 8.5x11 poster has a ratio of 0.016. It may be possible to use these ratios as a heuristic to estimate optimal physical text size or optimal distance for a picture of a poster.

Evaluating Repeated Measures: In this experiment, we took several pictures of the same poster from the same distance (2 feet away). Our goal for this experiment is to determine how often results are repeated. We took five pictures of the same 8.5x11 poster where the poster was fully in frame. There were a total of 18 RSIDs contained in the poster, although only 10 are examined here.

Table 7 – Results for repeated measure of the same poster

RSID	Frequency of summarization
Rs77803164	5
Rs72960926	5
Rs6931514	5
Rs4778720	5
Rs4680534	5
Rs401681	5
Rs3857067	5
Rs2271749	5
Rs166040	5
Rs149183310	5
Rs1368011	5
Rs12457997	5
Rs7904985	4
Rs33980500	4

Rs1808470	4
Rs1057910	4
Rs55874825	3
Rs12210050	3

As Table 7 shows, 12 of 18 RSIDs were recognized every time, with the remaining 6 being recognized between 3 and 4 times of the 5. No RSIDs in the poster were misrecognized. After assessing the location of the RSIDs that occurred at a frequency of 3 and 4, we could not determine any systematic source to account for the imperfect performance.

4 Extensibility

EXPLANE can be extended in a number of ways. As seen in Figure 1, the interface organizes the annotation results by categories (e.g. variant information, URLs, etc.). Adding or customizing tabs is the easiest way to use and extend the framework (full implementation instructions accompanied with code). Future work may include extending EXPLANE to include the BioWAP resource, as this would connect the OCR capability with additional protein-related resources that may provide a richer user experience. Furthermore, if the error rate of named entity recognitions is deemed tolerable, EXPLANE may be extended to recognize genes and query MyGene.info, providing a wider range of recognizable words with attached information. We anticipate that named entity recognition for gene names from scientific texts will improve over time, and that EXPLANE may eventually be updated to also process gene names. In addition to improvements in NER, performance of EXPLANE may conceivably be improved through training the Tesseract software on a scientific corpus (akin to the type of posters the OCR system will be processing), and through providing Tesseract a data dictionary containing domain-specific vocabulary terms that have been historically difficult to process, such as human gene names.

The application works very well under simulated conditions, but begins to have imperfect performance when pictures are taken of physical posters. From this observation, we infer that performance may vary depending on the characteristics and features of the mobile device being used. In certain cases, this may mean taking a picture in “landscape” mode rather than “portrait”. An ideal poster image would minimized angles (i.e. taking a picture in as centered an

orientation as possible), maximize text size (i.e. being as close to the poster as possible), ensure adequate lighting (for exposure), and minimize blurs.

Overall, the capability of EXPLANE is bounded by OCR recognition performance, and the content contained in MyVariant.info. However, as mentioned above, the OCR service we have deployed is not optimized, and MyVariant.info is ever-growing in terms of supported biological resources.

5 Conclusion

EXPLANE provides two contributions. First, it provides an extensible framework for poster annotation that any researcher can extend, or build upon. Second, it enables readers of posters to enrich their experience with supplemental poster information in real-time. Future work may entail adding social media capability, data downloads, emailing results, and user annotation. EXPLANE is readily extensible by anyone interested in downloading the source code and standing up their own instance. Performance may be improved by optimizing Tesseract to recognize RSIDs.

Funding: This work was supported by the US National Institute of Health grants U01HG008473. This publication was supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UL1 TR000423. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest: none declared.

References

- [1] B. J. Fogg, *Persuasive Technology: Using Computers to Change What We Think and Do*, vol. 5, no. 1. 2003.
- [2] S. Kumar, K. Boccia, M. McCutchan, and J. Ye, "Exploring spatial patterns of gene expression from Fruit Fly embryogenesis on the iPhone," *Bioinformatics*, vol. 28, no. 21, pp. 2847–2848, 2012.
- [3] P.-V. Nguyen, C. S. Verma, and S. K.-E. Gan, "DNAApp: a mobile application for

- sequencing data analysis.,” *Bioinformatics*, vol. 30, no. 22, pp. 3270–1, 2014.
- [4] C. Speake, S. Presnell, K. Domico, B. Zeitner, A. Bjork, D. Anderson, M. J. Mason, E. Whalen, O. Vargas, D. Popov, D. Rinchai, N. J. Chiche, L. Chiche, C. Quinn, and D. Chaussabel, “An interactive web application for the dissemination of human systems immunology data,” *J. Transl. Med.*, pp. 1–10, 2015.
- [5] M. L. de Oliveira and E. Galembeck, “Mobile Applications in Cell Biology Present New Approaches for Cell Modelling,” *J. Biol. Educ.*, vol. 9266, no. April, pp. 1–14, 2015.
- [6] C. P. B. Yiu and Y. W. Chen, “High-quality macromolecular graphics on mobile devices: A quick starter’s guide,” *Methods Mol. Biol.*, vol. 1091, pp. 343–352, 2014.
- [7] G. B. Quinn, C. Bi, C. H. Christie, K. Pang, A. Prli??, T. Nakane, C. Zardecki, M. Voigt, H. M. Berman, P. E. Bourne, and P. W. Rose, “RCSB PDB Mobile: IOS and Android mobile apps to provide data access and visualization to the RCSB Protein Data Bank,” *Bioinformatics*, vol. 31, no. 1, pp. 126–127, 2015.
- [8] S. V Eleftheriou, M. M. Bourdakou, E. I. Athanasiadis, and G. M. Spyrou, “GeneStoryTeller: a mobile app for quick and comprehensive information retrieval of human genes,” *Database (Oxford)*, vol. 2015, pp. 1–5, 2015.
- [9] J. Hakenberg and U. Leser, “What makes a gene name? Named entity recognition in the biomedical literature,” *Brief. Bioinform.*, vol. 6, no. 4, pp. 357–369, 2005.
- [10] J. Wermter, K. Tomanek, and U. Hahn, “High-performance gene name normalization with GeNo,” *Bioinformatics*, vol. 25, no. 6, pp. 815–821, 2009.
- [11] J. Hakenberg, S. Bickel, C. Plake, U. Brefeld, H. Zahn, L. Faulstich, U. Leser, and T. Scheffer, “Systematic feature evaluation for gene name recognition.,” *BMC Bioinformatics*, vol. 6 Suppl 1, p. S9, 2005.
- [12] C. Blaschke, E. A. Leon, M. Krallinger, and A. Valencia, “Evaluation of BioCreAtIvE assessment of task 2.,” *BMC Bioinformatics*, vol. 6 Suppl 1, p. S16, 2005.
- [13] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck, “ProMiner: rule-based protein and gene entity recognition.,” *BMC Bioinformatics*, vol. 6 Suppl 1, p. S14, 2005.
- [14] T. Mitsumori, S. Fation, M. Murata, K. Doi, and H. Doi, “Gene/protein name recognition based on support vector machine using dictionary as features.,” *BMC Bioinformatics*, vol. 6 Suppl 1, p. S8, 2005.
- [15] M. Torii, Z. Hu, C. H. Wu, and H. Liu, “BioTagger-GM: A Gene/Protein Name Recognition

- System,” *J. Am. Med. Informatics Assoc.*, vol. 16, no. 2, pp. 247–255, 2009.
- [16] D. M. Lowe and R. A. Sayle, “LeadMine: A grammar and dictionary driven approach to entity recognition,” *J. Cheminform.*, vol. 7, no. Suppl 1, p. S5, 2015.
- [17] R. Ding, C. N. Arighi, J. Y. Lee, C. H. Wu, and K. Vijay-Shanker, “pGenN, a gene normalization tool for plant genes and proteins in scientific literature,” *PLoS One*, vol. 10, no. 8, pp. 1–24, 2015.
- [18] J. Xin, A. Mark, C. Afrasiabi, G. Tsueng, M. Juchler, N. Gopal, G. S. Stupp, T. E. Putman, B. J. Ainscough, O. L. Griffith, P. L. Whetzel, C. J. Mungall, S. D. Mooney, and I. Andrew, “MyGene.info and MyVariant.info: Gene and Variant Annotation Query Services Jiwen,” 2015.
- [19] P. Riikonen, J. Boberg, T. Salakoski, and M. Vihinen, “Mobile access to biological databases on the Internet,” *IEEE Trans. Biomed. Eng.*, vol. 49, no. 12 II, pp. 1477–1479, 2002.
- [20] P. Riikonen, J. Boberg, T. Salakoski, and M. Vihinen, “BioWAP, mobile Internet service for bioinformatics,” *Bioinformatics*, vol. 17, no. 9, pp. 855–6, 2001.
- [21] M. Bostock, V. Ogievetsky, and J. Heer, “D³: Data-Driven Documents,” *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2301–9, Dec. 2011.
- [22] R. Smith, “An overview of the tesseract OCR engine,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 2, pp. 629–633, 2007.
- [23] “Node.js.” [Online]. Available: <https://nodejs.org/en/>.
- [24] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent, “The UCSC Table Browser data retrieval tool,” *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D493–6, 2004.