MADCAT Industry Day

- Joseph Olive joseph.olive@darpa.mil

Announcement

DARPA seeks strong, responsive proposals from well-qualified sources for a new language technology program called Multilingual Automatic Document Classification Analysis and Translation (MADCAT). The goal of this program is to automatically convert foreign language text images into English transcripts, thus eliminating the need for linguists and analysts while automatically providing relevant, distilled actionable information to military command and personnel in a timely fashion.

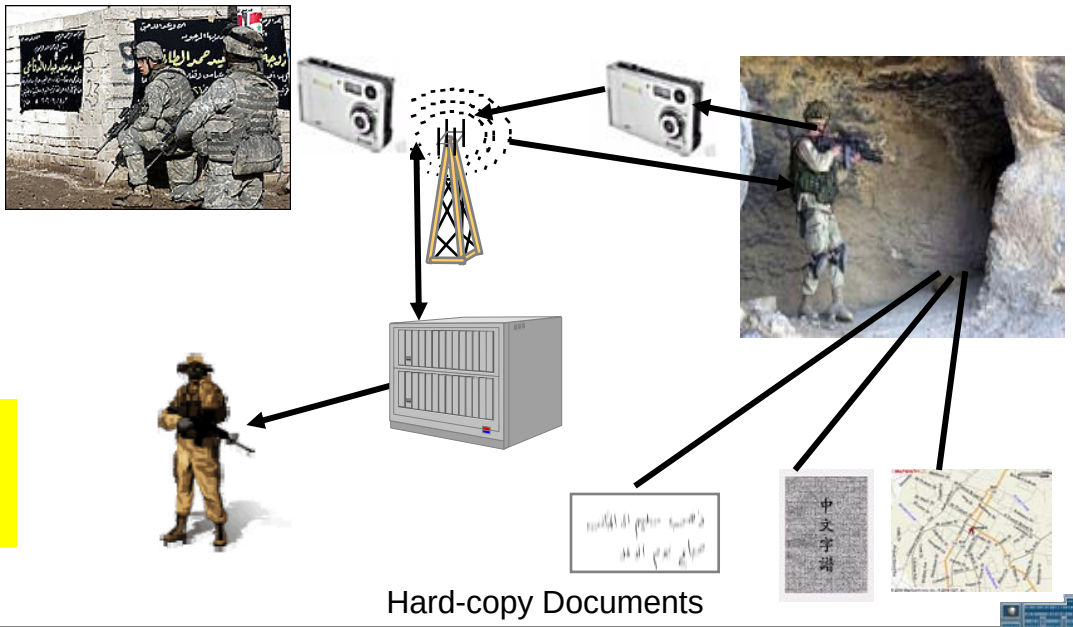
Goal

Provide soldiers in the field
relevant, distilled, actionable information
automatically in a timely fashion
from all printed or hand-written Sources
in multiple languages

- Strategic Plans, maps
- Financial Records
- Graffiti, propaganda
- Manuals

⋮

Situational awareness
In urban warfare



Why is this technology important?

Is this graffiti or is it important?

Necessary for situation awareness – especially for urban warfare

Example Input - Printed

قناة MBC

٧,١٥ ديجيمون	٦,٠٠ نشرة الأخبار
٨,١٥ جونى كويست	٦,٣٠ من سيربح المليون
٨,٥٠ أجواء وأطباق	٨,٠٠ لا تفهمونا غلط
٩,١٥ صلاة الجمعة ـ مباشر من مكة المكرمة	٩,٣٠ مشاهد مثيرة
١٠,٠٥ الإفتاء ـ مباشر	١٠,٠٠ الأسبوع السياسى
١١,٣٠ فيلم عربى	١١,٠٠ مسلسل عربى
١,٠٠ الأخبار	١٢,٣٠ صلالة ٢٠٠١
١,٣٠ برنامج رياضى	١٢,٣٥ بيبسى كورة
١,٥٥ الحياة البرية	١,٠٠ صلالة ٢٠٠١
٣,٠٠ موجز الأنباء	٢,٠٠ فيلم أجنبى
٣,٠٥ دورى مستر كارد	٣,٠٠ ديجيمون
٣,٣٥ حوار الأسبوع	٣,٢٥ تيمون وبومبا
٤,٠٠ موجز الأنباء	٣,٥٠ مهووس فيكى
٤,٠٥ مفكرة المراسل	٥,٠٠ عيون الحب
٤,٣٠ الأسبوع السياسى	٥,٤٥ منوعات غنائية
	٦,٠٠ مسلسل عربى

قناة المستقبل

٤,٠٠ كارتون	١٣,٣٠ نادى ديزنى
٧,٠٠ عالم الصباح	١٥,٠٠ كمبس
٩,٠٠ صلاة الجمعة	١٦,٠٠ شبكة
١٠,٠٠ المطبخ	١٧,٠٠ الأخبار
١١,٠٠ موجز الأخبار	١٧,٣٠ ميشو شو
١١,٠٥ فقرة إخبارية	١٨,٣٠ فيلم أجنبى
١١,٣٠ فانيست هوم فيديو	٢٠,٣٠ نص دقيقة
١٢,٠٠ الأخبار	٢١,٣٠ فيلم عربى
١٣,٠٠ نشرة الأخبار	٢٣,٣٠ مسلسل عربى

وسهير المرشدى ومديحة حمدى وعبد الرحمن أبو زهرة واخراج توفيق صالح ويدور حول جماعة من الصيادين يفاجأون بزميلهم عبد الموجود وقد اشترى مركبا كبيرا يعمل بالبخار وبذلك يستولى على النصيب الأكبر من الصيد، ويشير عليهم ابن السيد البلطى أقدم الصيادين بالعمل معا كشركة للصيد الا أنهم يتمكنوا من شراء مركب ينافسون به عبد الموجود.

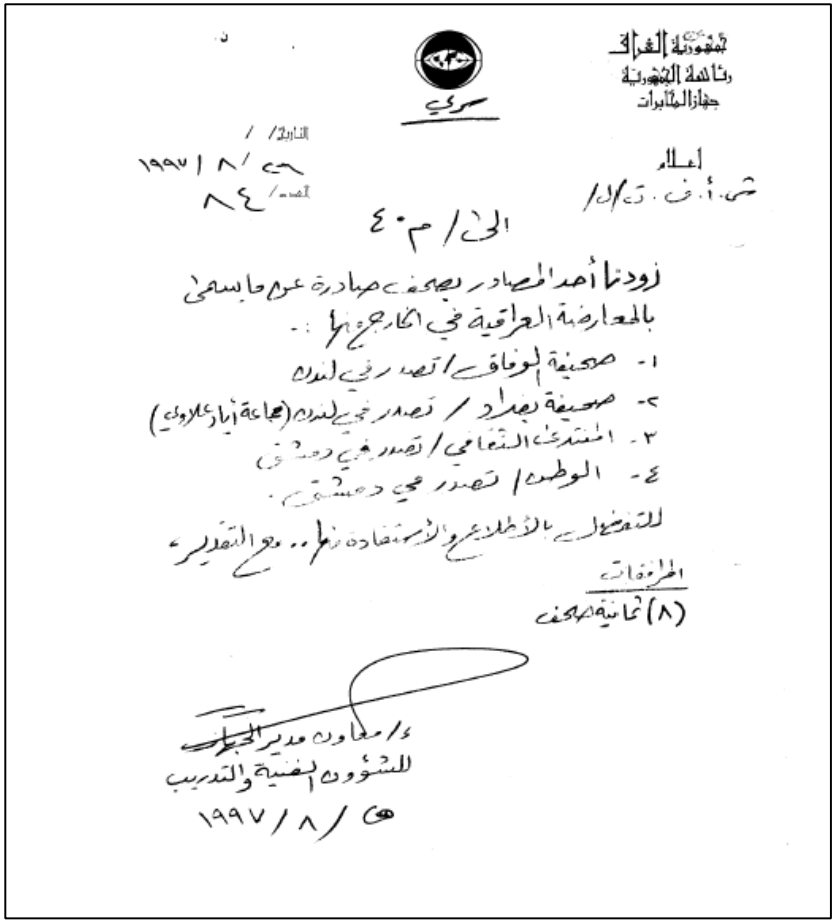
..و«الثلاثة يحبونها» و«الثأر»
فيلمان عربيان على الفضائية الثانية

محمود يس

● تعرض القناة الفضائية الثانية الساعة ٢,٠٠ ظهرا بتوقيت القاهرة فيلم (الثلاثة يحبونها) بطولة سعاد نصر وحسن يوسف ويوسف شعبان ويوسف فخر الدين وناهد شريف واخراج محمود ذو الفقار.وتدور أحداثه حول الفتاة (ايمان) الموظفة باحدى الشركات والتى تسىء فهم الحرية وتعيش ثلاث قصص فى وقت واحد مع (عصام) مديرها والذى يطلبها للزواج لكنها تصمم على مواصلة دراستها الجامعية، و(كمال) الشاب خفيف الظل و(عادل) الذى ينتقد تصرفاتها المستهترة

● وتعرض الساعة ١,٠٠ بعد منتصف الليل بتوقيت القاهرة فيلم (الثأر) بطولة يسرا ومحمود يس وحمدى الوزير وفاروق يوسف واخراج محمد خان وفيه يقوم أربعة أشخاص باختطاف ندا من زوجها أحمد الذى يلتقط رقم السيارة ويبلغ الشرطة، ويصمم على الانتقام منهم بنفسه ويقوم بقتل ثلاثة أشخاص ويتم القبض عليه قبل ارتكابه الجريمة الرابعة.

Example Input - Handwritten



جمهورية العراق
رئاسة الجمهورية
جهاز المخابرات

سري

التاريخ / /
٢٥ / ٨ / ١٩٩٧
العدد / ٨٤

اعلام
ش.أ.ف.ت/١/

الى / م٤٠

زودنا أحد المصادر بصحف صادرة عن ما يسمى
بالمعارضة العراقية في الخارج وهي :-
١- صحيفة الوفاق / تصدر في لندن
٢- صحيفة بغداد / تصدر في لندن (جماعة اياد علاوي)
٣- المنتدى التقدمي / تصدر في دمشق
٤- الوطن / تصدر في دمشق .
للتفضل بالاطلاع والاستفادة منها .. مع التقدير

المرفقات
(٨) ثمانية صحف

ع/ معاون مدير الجهاز
للشؤون الفنية والتدريب
٢٥ / ٨ / ١٩٩٧

Example Input - Mixed

بسم الله الرحمن الرحيم

امانة سر الحرس الجمهوري

سري وشخصي

رئاسة الجمهورية / المكتب الخاص

السكرتير :

م/ الاسرى الكويتيين

تنفيذاً لأمر السيد الرئيس {القائد صدام حسين} {حفظة الله ورعاة}

حسب ما جاء بقرار مجلس قيادة الثورة في يوم الجمعة

بتاريخ {٤ / ٣ / ٢٠٠٣}.

بنقل كافة الأسرى الكويتيين / البالغين عددهم {٤٤٨}. اسيراً كويتيين .

والمودعين / في سجن {الفداء الاغر}

أ المخابرات / المركز العام وسجن كاظمة ب {الكاظمة}.

لجعلهم دروع بشرية في كافة المواقع التي يتوقع ضربها من قبل الامريكان

المعتدين / من مواقع الاتصالات والوزارات الاساسية / الاذاعة والتلفزيون /

منشئات التصنيع العسكري /

وكافة المواقع يتوقع الاعتداء عليها من قبل العدوانيين المجرمين الانجلو

امريكان

تناط مهمة نقلهم بالتنسيق بين كل من :

مديرية جهاز المخابرات

رئاسة اركان الحرس الجمهوري

وبأشراف مباشر من قبل جهاز الامن الخاص / امن الجهاز

قصي صدام حسين

المشرف

على امانة سر الحرس الجمهوري

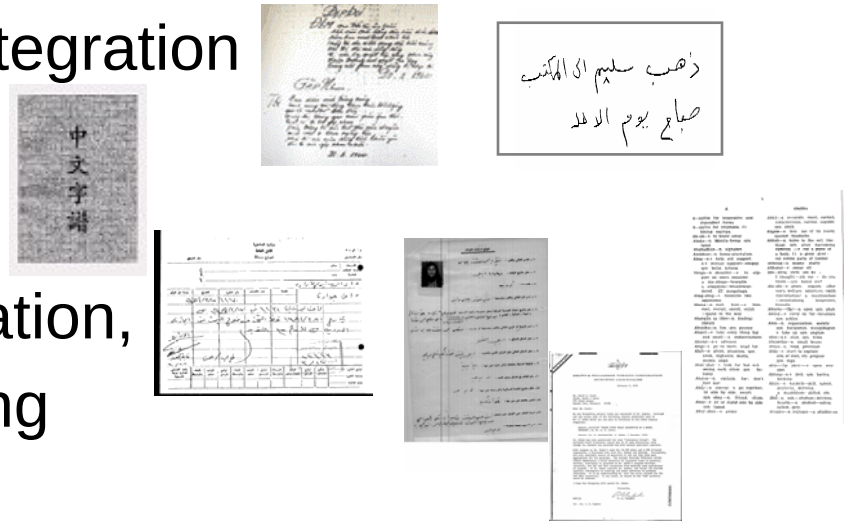
صورة منه الى :

مديرية جهاز المخابرات / مكتب مدير الجهاز

رئاسة اركان الحرس الجمهوري / مكتب رئيس الأركان

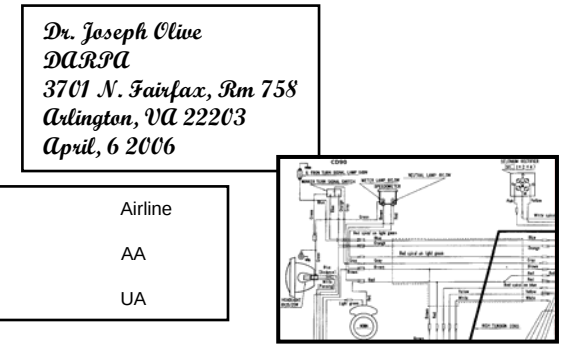
Technology Needs

- Create a framework for integration
- Language/Script ID
- Document Type Classification, Segmentation and Labeling



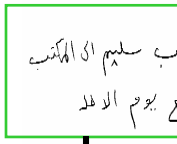
- Zone Analysis and Interpretation
- Robust Multilingual OCR
- Machine Translation
- Distillation

GALE



Deliverables

- Translation
- Classification
  - Author
  - Recipient
  - Names, entities, events content
  - Topics
  - Logos
  - Language ID



MADCAT SYSTEM

Translation

Classification:

Language ID, Document Type, Author, Recipient, Logo, Date, Subject, Relevance, Content

Example Output - Printed

قناة MBC

٧,١٥ ديجيمون	٦,٠٠ نشرة الأخبار
٨,١٥ جونى كويست	٦,٣٠ من سيربح المليون
٨,٥٠ أجواء وأطباق	٨,٠٠ لا تفهمونا غلط
٩,١٥ صلاة الجمعة ـ مباشر من مكة المكرمة	٩,٣٠ مشاهد مثيرة
١٠,٠٥ الإفتاء ـ مباشر	١٠,٠٠ الأسبوع السياسى
١١,٣٠ فيلم عربى	١١,٠٠ مسلسل عربى
١,٠٠ الأخبار	١٢,٣٠ صلالة ٢٠٠١
١,٣٠ برنامج رياضى	١٢,٣٥ بيبسى كورة
١,٥٥ الحياة البرية	١,٠٠ صلالة ٢٠٠١
٣,٠٠ موجز الأنباء	٢,٠٠ فيلم أجنبى
٣,٠٥ دورى مستر كارد	٣,٠٠ ديجيمون
٣,٣٥ حوار الأسبوع	٣,٢٥ تيمون وبومبا
٤,٠٠ موجز الأنباء	٣,٥٠ مهووس فيكى
٤,٠٥ مفكرة المراسل	٥,٠٠ عيون الحب
٤,٣٠ الأسبوع السياسى	٥,٤٥ منوعات غنائية
	٦,٠٠ مسلسل عربى

قناة المستقبل

٤,٠٠ كارتون	١٣,٣٠ نادى ديزنى
٧,٠٠ عالم الصباح	١٥,٠٠ كمبس
٩,٠٠ صلاة الجمعة	١٦,٠٠ شبكة
١٠,٠٠ المطبخ	١٧,٠٠ الأخبار
١١,٠٠ موجز الأخبار	١٧,٣٠ ميشو شو
١١,٠٥ فقرة إخبارية	١٨,٣٠ فيلم أجنبى
١١,٣٠ فانيست هوم فيديو	٢٠,٣٠ نص دقيقة
١٢,٠٠ الأخبار	٢١,٣٠ فيلم عربى
١٣,٠٠ نشرة الأخبار	٢٣,٣٠ مسلسل عربى

وسهير المرشدى ومديحة حمدى وعبد الرحمن أبو زهرة واخراج توفيق صالح ويدور حول جماعة من الصيادين يفاجأون بزميلهم عبد الموجود وقد اشترى مركبا كبيرا يعمل بالبخار وبذلك يستولى على النصيب الأكبر من الصيد، ويشير عليهم ابن السيد البلطى أقدم الصيادين بالعمل معا كشركة للصيد ألا انهم يتمكنوا من شراء مركب ينافسون به عبد الموجود.

..و«الثلاثة يحبونها» و«الثأر»

فيلمان عربيان على الفضائية الثانية

● تعرض القناة الفضائية الثانية الساعة ٢,٠٠ ظهرا بتوقيت القاهرة فيلم (الثلاثة يحبونها) بطولة سعاد نصر وحسن يوسف ويوسف شعبان ويوسف فخر الدين وناهد شريف واخراج محمود ذو الفقار.وتدور أحداثه حول الفتاة (ايمان) الموظفة باحدى الشركات والتى تسىء فهم الحرية وتعيش ثلاث قصص فى وقت واحد مع (عصام) مديرها والذى يطلبها للزواج لكنها تصمم على مواصلة دراستها الجامعية، و(كمال) الشاب خفيف الظل و(عادل) الذى ينتقد تصرفاتها المستهترة

● وتعرض الساعة ١,٠٠ بعد منتصف الليل بتوقيت القاهرة فيلم (الثأر) بطولة يسرا ومحمود يس وحمدى الوزير وفاروق يوسف واخراج محمد خان وفيه يقوم أربعة أشخاص باختطاف ندا من زوجها أحمد الذى يلتقط رقم السيارة ويبلغ الشرطة، ويصمم على الانتقام منهم بنفسه ويقوم بقتل ثلاثة أشخاص ويتم القبض عليه قبل ارتكابه الجريمة الرابعة.

محمود يس

English

Table

Logo

Reverse Contrast

Two-column text (Arabic)

Text Headline (Arabic)

Graphic (Face)

Caption (Name)

Single column text

Channel MBC

57:1 Djimon

8:15 Johnny Quest

8:50 Points of Views

9:15 Friday Prayer – Live from Mecca Al-Moukrama

10:05 Religious Decree - Live

1:00 News

1:30 Sports Program

6:00 Who wants to be a millionaire

8:00 Don't misunderstand us

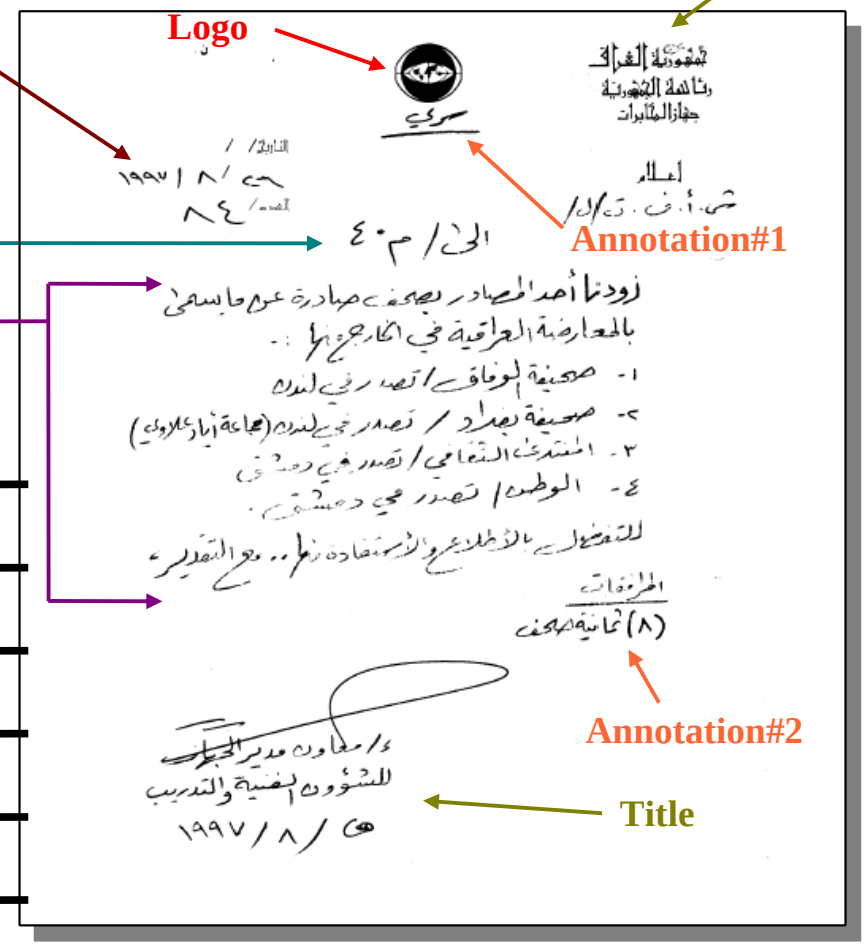
And Suheer Al-Murshdi and Madeehah Hamdi and Abed Al-Rahman Abu Zahra and directed by Tawfeeq Saleh it is about a group of fishermen surprised by their mate Abed Al-Mawjoud when he buys a big boat that works on steam and in return he gains the bigger share of the fishing. Ibin Al-Sayed Al-Balti, who is the oldest fishermen, suggests to them to work together as a fishing company so that they can buy a boat to compete with Abed Al-Mawjoud.

"The Three Love Her" and

"The Retaliation"

two Arabic films on the Al-Fathaeia Al-Thania (Satellite 2)

Example Output - Handwritten




```
<xml>
<meta ID="BIAP-2003-000090.pdf" genre="correspondence"
      language="Arabic" >
<date obtained="Jun-10-2005" created="Aug-26-1997">
<author title="Assistant manager training and technical
      Department" address="The Republic of Iraq, President's
      Office, Intelligence Department">
<recipient name="m40">
<annotation#1 translation="Confidential">
<text translation="One of the sources informed us of newspapers
      that are so called the Iraqi defiance abroad, some of them
      are:"
      <list><item 1-Al Wafaa newspaper published in London>
      <item 2-Baghdad newspaper published in London (Ayad
      Alawi group).>
      <item 3-Al Muftada al Thakafi published in Damascus>
      <item 4-Al Wattan published in Damascus></list>
      "for your information.  Regards">
<annotation#2 translation="Attachments 8 newspapers">
</xml>
```

Example Output - Mixed

```
<xml>
<meta ID="CMPC-2003-012666.tif" genre="correspondence"
      language="Arabic" >
<date obtained="Mar-22-2006" created="Mar-14-2003">
<author name="Qusai Saddam Hussein"
      title="Supervisor of the Republican Guard Secretariat">
<recipient name="Presidential Office/ Special Office">
...
<text translation="Regarding the execution of Mr. President,
      Commander Saddam Hussein's (God protect him) orders,
      according to the decision of the Revolutionary Command
      Council on Friday, March 4, 2003.

      Transfer all Kuwaiti POW's / a total of 448 captured Kuwaitis
      who are located at the Al-Nida Al-Agher Prison and the
      Intelligence / General Center and Kazema Prison in Al-
      Kazema, to make them human shields at all locations that
      are expected to be attacked by the American aggressors.">
...
</xml>
```

Program Requirements

- Four Tasks
  - Printed documents
  - Handwritten and mixed documents
  - Data
  - Evaluation
- Arabic Only
  - Surprise language in later phases
- Only English output evaluated

Evaluation paradigms

- Evaluating Full Translation
  - A set of documents in Arabic (20K words per language)
  - Create a “gold standard”
  - Human editing to match meaning of machine output and “gold standard”
  - Accuracy is 1-(no. of edits)/(no. of words)
- Evaluating Metadata (Logos, Authors, Recipient, Date)
  - A set of documents not seen by the contractors (10K pages)
  - Manually establish ground truth
  - Machine encoded metadata
  - Establish ROC curves and EER points

Metrics

	Translation from Machine Print		Translation from Handwriting		Classification
Phase	% Accuracy	% of documents	% Accuracy	% of documents	% Accuracy
Baseline	65%	70%	2%	50%	60%
Phase 1	75%	80%	40%	70%	75%
Phase 2	85%	85%	60%	80%	85%
Phase 3	90%	85%	75%	85%	95%
Phase 4	95%	90%	85%	90%	Completed
Phase 5	95%	95%	90%	90%	Completed