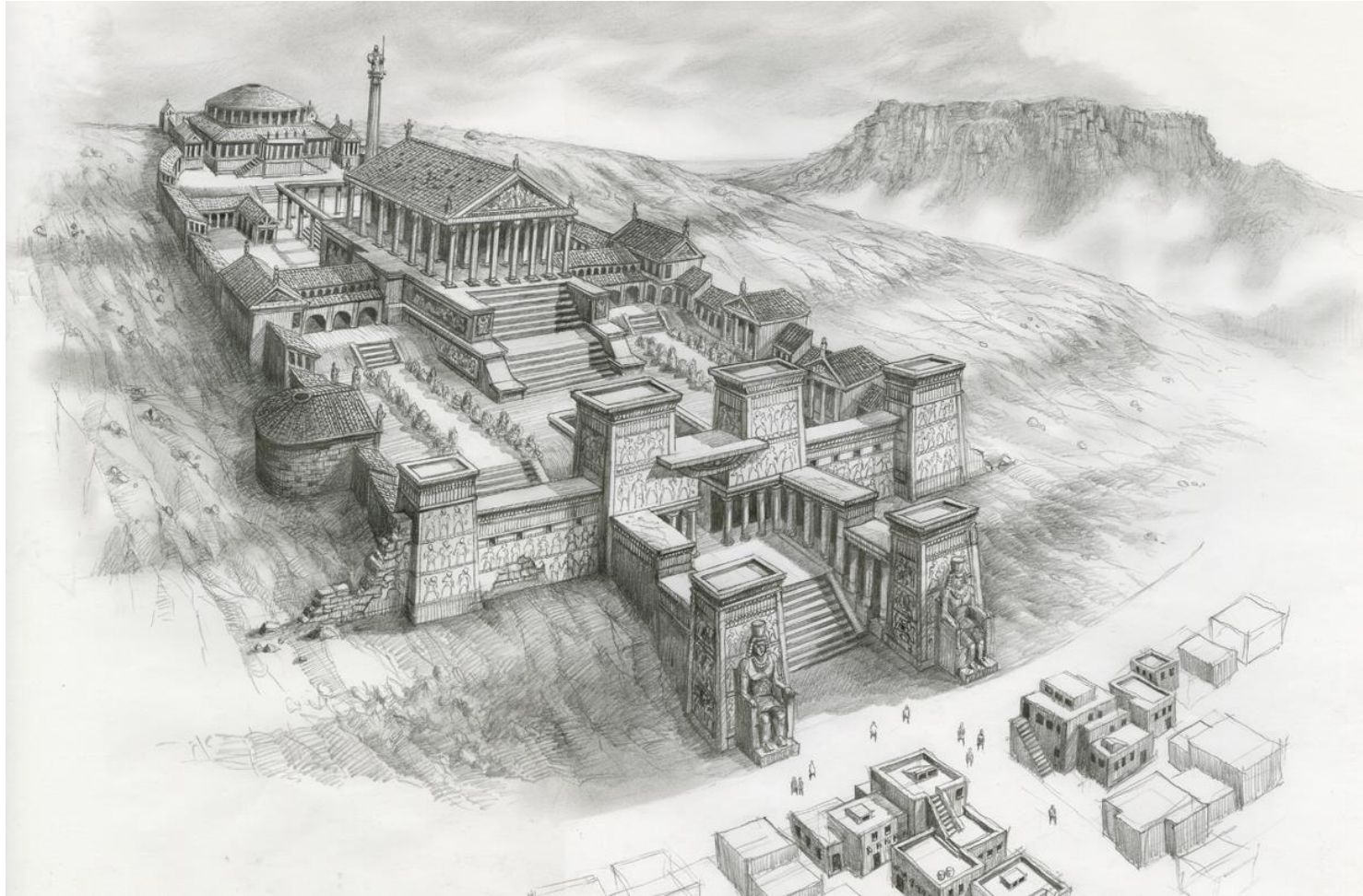# PROJECT FACTOR

Speech to Text – innovating and integrating WITH Data

LIBRARY OF ALEXANDRIA

WHY?

Libraries were the controlled center of knowledge for the entire history of mankind!

Until the digital age…

Currently the online services groups in universities have become the defacto center for knowledge

But to them – the data they control or that knowledge is a burden to store and archive

They throw it away every 1 to 1.5 years on average!

Wouldn't it be great if we have a place and a group that were experts in categorizing, searching, retrieving, and presenting this data to those who are looking for it?

WE DO…

It is called the library…

PROJECT FACTOR PROPOSAL

Factor is a software framework

- The main goal is to take text from speech (video or audio) and convert it to raw text that users can manage and mark up

Why bother?

- Users can correct the failures of machine translation
- Users can then provide full English transcripts for subtitles and ADA use
- Users can mark up unstructured text and share it with others in a library
- Users can search the gathered text and make custom books from content scattered across departments and even years of lectures.
- Users can also search via keywords to find video as well as text

WHAT ARE THE COMPONENTS?

Speech to text engine – using an opensource library called Sphinx from Carnegie Mellon University

Project Convert

- Note… they are already using this library for a transcript purpose for some of their own courses
- Transcript also provides timestamp which is used in the next project

Project Window-Pane

- This is the interface that allows the user to view video (similar to blackboard) and simultaneously edit the text transcript

Project Marker

- This portion allows a user to enter a drag and drop application and mark unstructured text via a simple GUI interface.
- Code is rendered into a DocBook or HTMLBook format which allows for export to PDF, HTML, Docbook, and even ePub
- PDF and HTML formats can be printed too – making real-time books

FACTOR COMPONENTS

Project SH-3

- This is a meta portion of the project that deals with raw data storage and preparing data for archiving and retrieval.

Project Promised-Land

- This is the search engine portion
- Aggregates all the searchable text transcripts
- Uses Lucene (an opensource indexing tool that prepares data for searching)
- Uses Solr (an opensource webcrawler that a user would search from)
- Uses Hadoop (just for data storage)

WHAT IS THE CURRENT STATE?

We have a working prototype of each part

- Main UI parts have been integrated

- Two student groups have worked on separate prototypes and produced research into the Sphinx module and its inner workings

- I am looking for support from any interested parties.

- IPRO is not an option as they are looking for Full-time research professors

- My department would not sponsor me to do an IPRO

- So this idea is out unless someone else can pick up the torch

WHAT IS THE LIBRARY'S ROLE?

Help with licenses

- Can all the content be placed under creative commons?

- If so which one?

- Will the university have an issue?

- Are there intellectual property issues?

Funding

- Are there research grants for a product like this? IMLS?
  Mellon foundation? ADA?

- Currently the NAD is *suing* MIT and Harvard for not providing captioning on their online course videos.

- How would one apply? Where would one look?

Way down the road

- Management and shepherding of the project?

QUESTIONS?

Questions and comments?

- [Resources and prototypes publically available here](#)