INF-558/CSCI-563
Building Knowledge Graphs
Spring 2021

Craig Knoblock and Jay Pujara
University of Southern California

Course Overview

Challenges in Building Knowledge Graphs

- Finding the data
- Extracting it from structured sources
- Extracting it from unstructured sources
- Cleaning/normalizing the data
- Aligning the data at the schema level
- Finding connections in the data
- Efficiently representing the data
- Efficiently accessing the data
- Visualizing and analyzing the data

Course Mechanics

Course Materials and Web Pages

- Google Drive folder:
  - http://bit.ly/dsci558-s21
  - All slides, homeworks, etc. will be posted on the shared Dropbox folder
- Blackboard – blackboard.usc.edu
  - Submit homeworks on Blackboard
- Blackboard discussion board
  - All questions should be posted
    - Expect response from TA in 24 hours, otherwise let Profs know
    - If you know the answer to a posted question, please try to provide helpful suggestions
    - But please don't post answers to homeworks!

Prerequisites & Recommended

- Prerequisite(s):
  - INF 551 or CSCI 585
  - INF 552 or CSCI 567

- Recommended Background: Experience programming in Python

Grading

- Homework: 25%
  - Must be turned in on time, but
  - 25% penalty if late one week
  - Nothing accepted after one week
- Quizzes: 20%
  - One per class, at the beginning of the class
    - Questions based on the lectures, readings, and homeworks from previous week
  - No make ups
  - We will drop the two lowest grades
- Final Exam: 15%
- Project: 40%

More on Grading

- This is a hard class, but you will learn a lot!
  - Principles and theory
    - Technical readings and lectures (quizzes, final exam)
  - Putting principles into practice
    - Homeworks and project!
- Grade distribution

94 - 100 = A	74 – 76.9 = C
90 – 93.9 = A-	70 – 73.9 = C-
87 – 89.9 = B+	67 – 69.9 = D+
84 – 86.9 = B	64 – 66.9 = D
80 – 83.9 = B-	60 – 63.9 = D-
77 – 79.9 = C+	Below 60 is an F

Readings

- Listed in the syllabus
  - You can read them online or print them
  - We may update them as the semester goes on and will let you know if we do
- Read all required readings before class
  - You will get more out of the lectures
  - We can ask you about them in the quizzes

Slides

- Available online by the start of lecture
- Not intended as a replacement for the lecture
- You can bring them to class

Working Together

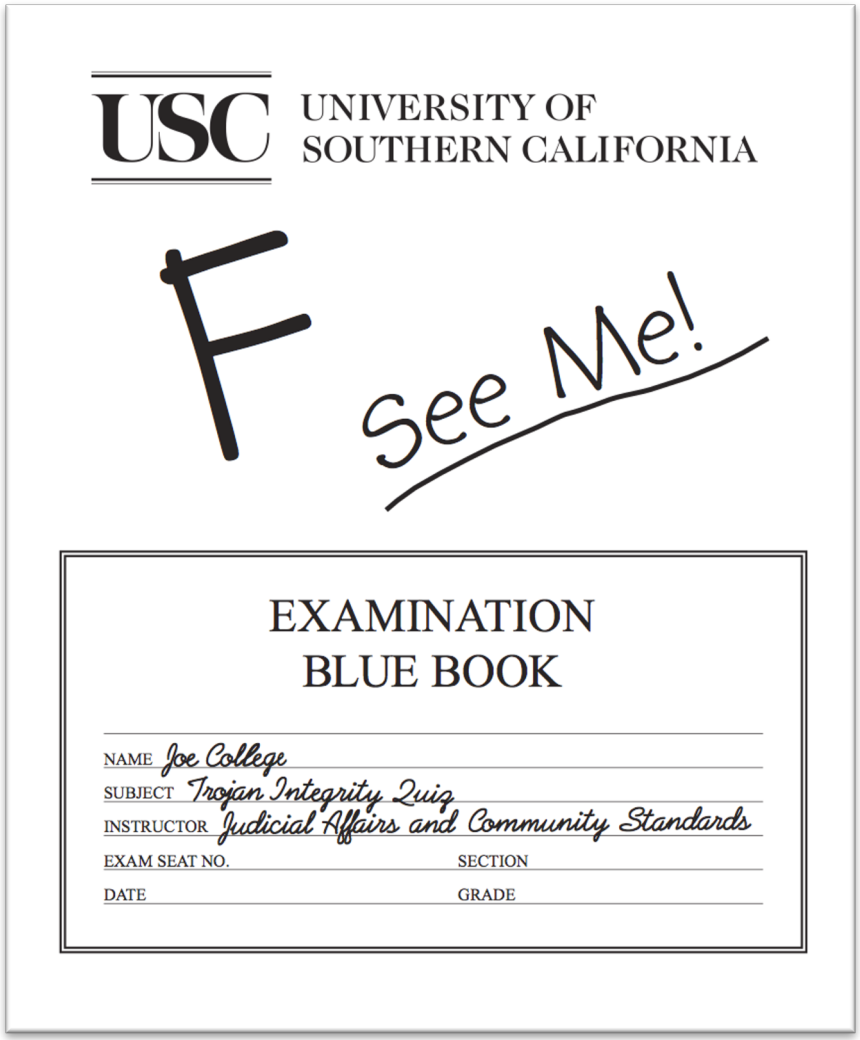
- Each person must do their own homework
  - We will check for overlap in homeworks
  - If we find any plagiarism, all parties lose credit and will be reported for cheating
    - Don't share your answers
    - Don't leave printouts in the trash with your answers
    - Don't give out your password
    - Don't copy others (they may have the wrong answer anyway!)
- You can ask the professors or TAs for help

Cheating

- Not tolerated!
- All infractions will be reported

- Examples:
  - Turning in someone else's homework
  - Doing the homework in collaboration with someone else and then turning in your own copy
  - Copying from someone else during a quiz or exam

We Follow USC Policies

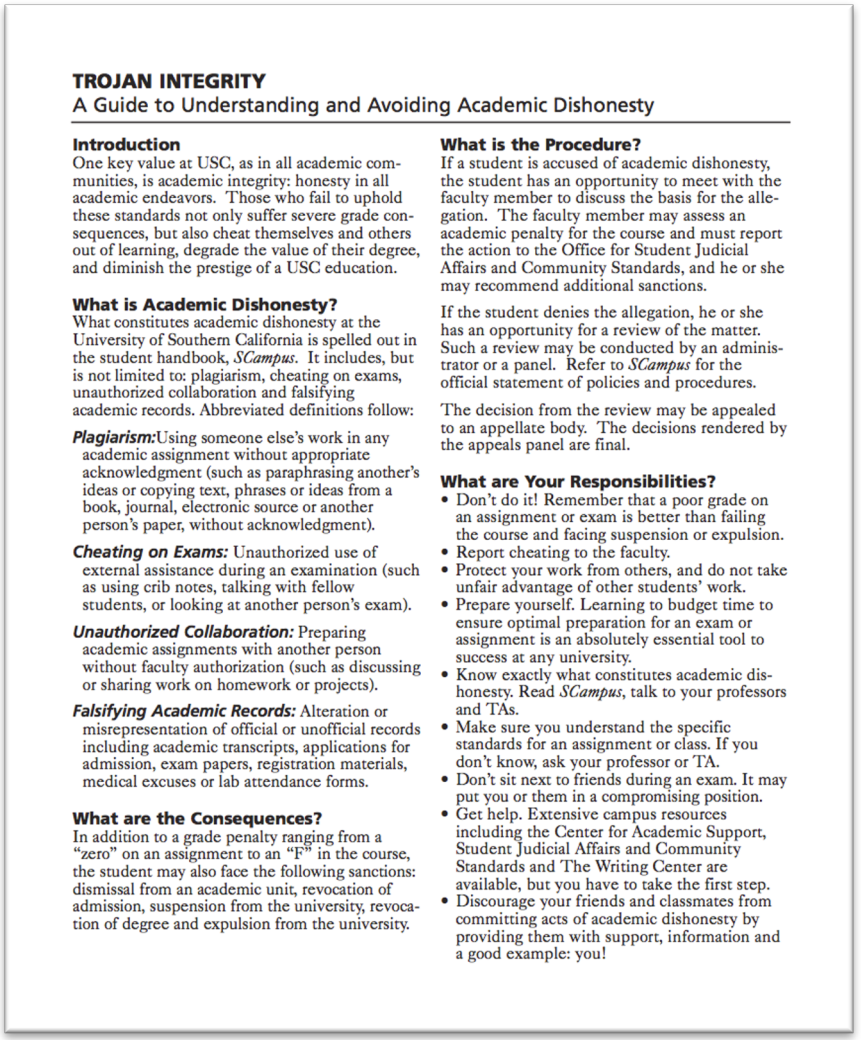


USC UNIVERSITY OF SOUTHERN CALIFORNIA

F See Me!

EXAMINATION
BLUE BOOK

NAME Joe College
SUBJECT Trojan Integrity Quiz
INSTRUCTOR Judicial Affairs and Community Standards
EXAM SEAT NO. SECTION
DATE GRADE



TROJAN INTEGRITY
A Guide to Understanding and Avoiding Academic Dishonesty

Introduction
One key value at USC, as in all academic communities, is academic integrity: honesty in all academic endeavors. Those who fail to uphold these standards not only suffer severe grade consequences, but also cheat themselves and others out of learning, degrade the value of their degree, and diminish the prestige of a USC education.

What is Academic Dishonesty?
What constitutes academic dishonesty at the University of Southern California is spelled out in the student handbook, *SCampus*. It includes, but is not limited to: plagiarism, cheating on exams, unauthorized collaboration and falsifying academic records. Abbreviated definitions follow:

Plagiarism: Using someone else's work in any academic assignment without appropriate acknowledgment (such as paraphrasing another's ideas or copying text, phrases or ideas from a book, journal, electronic source or another person's paper, without acknowledgment).

Cheating on Exams: Unauthorized use of external assistance during an examination (such as using crib notes, talking with fellow students, or looking at another person's exam).

Unauthorized Collaboration: Preparing academic assignments with another person without faculty authorization (such as discussing or sharing work on homework or projects).

Falsifying Academic Records: Alteration or misrepresentation of official or unofficial records including academic transcripts, applications for admission, exam papers, registration materials, medical excuses or lab attendance forms.

What are the Consequences?
In addition to a grade penalty ranging from a "zero" on an assignment to an "F" in the course, the student may also face the following sanctions: dismissal from an academic unit, revocation of admission, suspension from the university, revocation of degree and expulsion from the university.

What is the Procedure?
If a student is accused of academic dishonesty, the student has an opportunity to meet with the faculty member to discuss the basis for the allegation. The faculty member may assess an academic penalty for the course and must report the action to the Office for Student Judicial Affairs and Community Standards, and he or she may recommend additional sanctions.

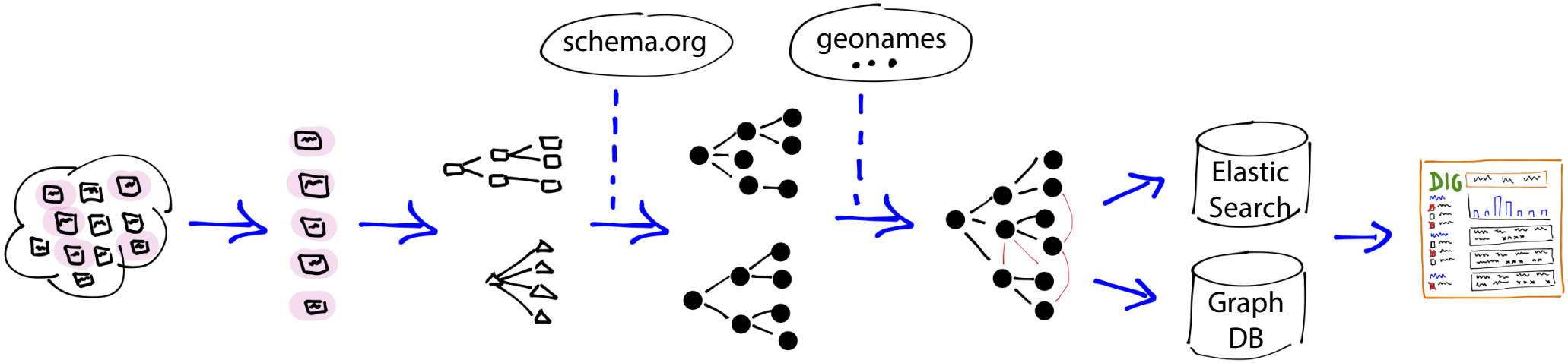
If the student denies the allegation, he or she has an opportunity for a review of the matter. Such a review may be conducted by an administrator or a panel. Refer to *SCampus* for the official statement of policies and procedures.

The decision from the review may be appealed to an appellate body. The decisions rendered by the appeals panel are final.

What are Your Responsibilities?
- Don't do it! Remember that a poor grade on an assignment or exam is better than failing the course and facing suspension or expulsion.
- Report cheating to the faculty.
- Protect your work from others, and do not take unfair advantage of other students' work.
- Prepare yourself. Learning to budget time to ensure optimal preparation for an exam or assignment is an absolutely essential tool to success at any university.
- Know exactly what constitutes academic dishonesty. Read *SCampus*, talk to your professors and TAs.
- Make sure you understand the specific standards for an assignment or class. If you don't know, ask your professor or TA.
- Don't sit next to friends during an exam. It may put you or them in a compromising position.
- Get help. Extensive campus resources including the Center for Academic Support, Student Judicial Affairs and Community Standards and The Writing Center are available, but you have to take the first step.
- Discourage your friends and classmates from committing acts of academic dishonesty by providing them with support, information and a good example: you!

Projects

Course Projects



Groups of Two

http://pedroszekely.blogspot.com/2013/07/tips-for-working-successfully-in-group.html

Grading Breakdown of the Project:

- Proposal: 10%
- Project video: 30%
- Presentation: 30%
- Overall project: 30%

Project Steps

1. Sales pitch (proposal)
2. Homeworks
3. More work
4. Presentation
5. Video

Project Pitch

2 minutes to pitch
3-5 minutes negotiation

Project Requirements

Solves an interesting problem

Has not been done

Is doable

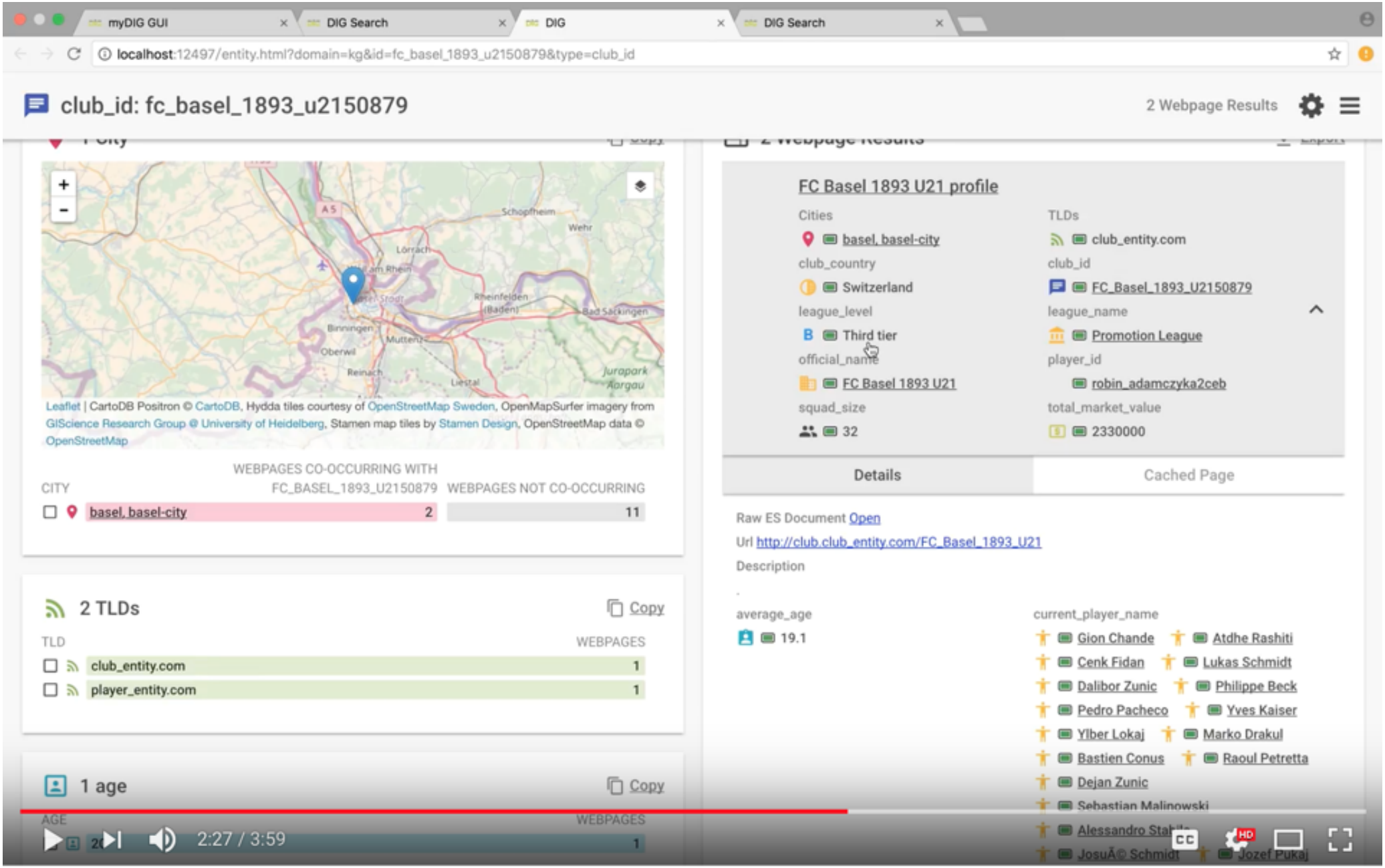
And…

Project Requirements (cont.)

- At least 3 sources total
- At least 1 structured
- At least 3 websites
- At least 20,000 pages total
- Smallest source at least 500 pages
- Structured sources at least 500 records
- Must solve at least 2 entity resolution problems
  - one per student
- Must define and populate at least 8 semantic types
- Must define at least one network analysis problem using the KG and solve it with standard libraries
- It is enough to use an existing GUI (e.g., DIG, NEO4J, ??)
  - No need to program a separate GUI

Example Project

https://www.youtube.com/watch?v=fDzKPhgmUDY



Final Project Presentation

Making presentations will be an essential part of my job

Yes

No

every group will present

PechaKucha 20x20

www.pechakucha.org

Szekely | Evernote | EverClip | The Milk | Blackboard | myUSC | myViterbi | DR Projects | G Groups | Tumblr | Blogger | G Analytics | Karma | WK Jira | OpenRdf | lod.isi | Fusion Aps

ATTEND | WATCH | CHANNELS

PechaKucha™ 20×20 IMAGES SECONDS

ABOUT | DAILY BLOG | Login | Sig

CONNECT WITH US: ON FACEBOOK ON TW

The art of concise presentations.

PechaKucha Night, now in over 700 cities, was devised in Tokyo in February 2003 as an event for young designers to meet, network, and show their work in public.

SEARCH

Display a menu

20 x 20 = 6:40

20 slides*

20 seconds/slide

auto-advance

* no builds, no transitions, no movies, no sounds

Good and Bad Advice on the Web

Tips for Success

1. …

2. The Rule of 1-6-6. One idea per slide, six bullets per slide, six words per bullet.

3. Use animations on your bullet points to control the amount of text on screen at once. When a full screen of text appears at once, students will tune out the teacher while reading. Better to set the pace of your presentation and introduce each bullet point when desired.

SDSU Department of Educational Technology

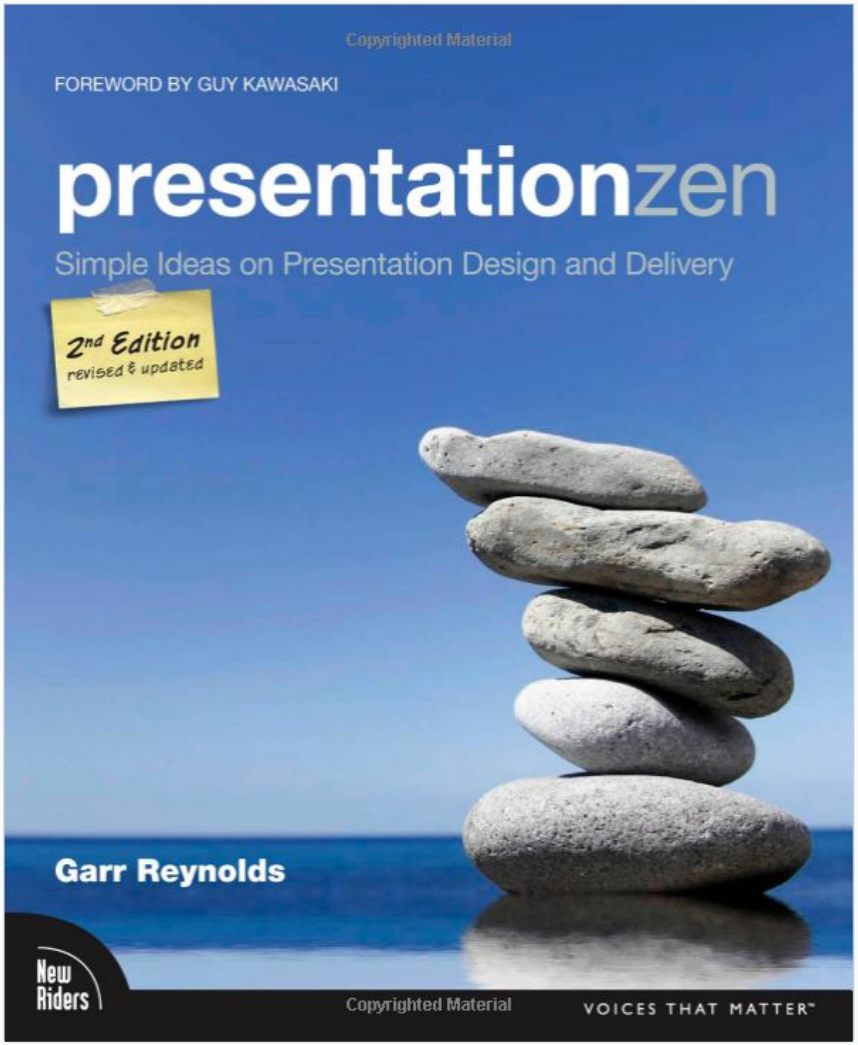
Presentation Guidelines

ONE slide of only bullets

=

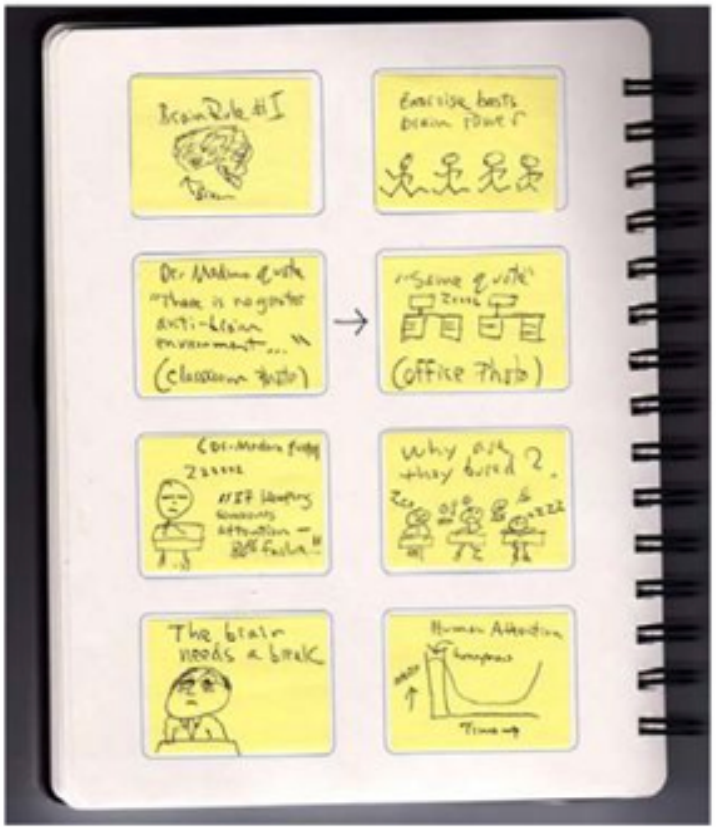
ZERO points in your presentation

Good Advice

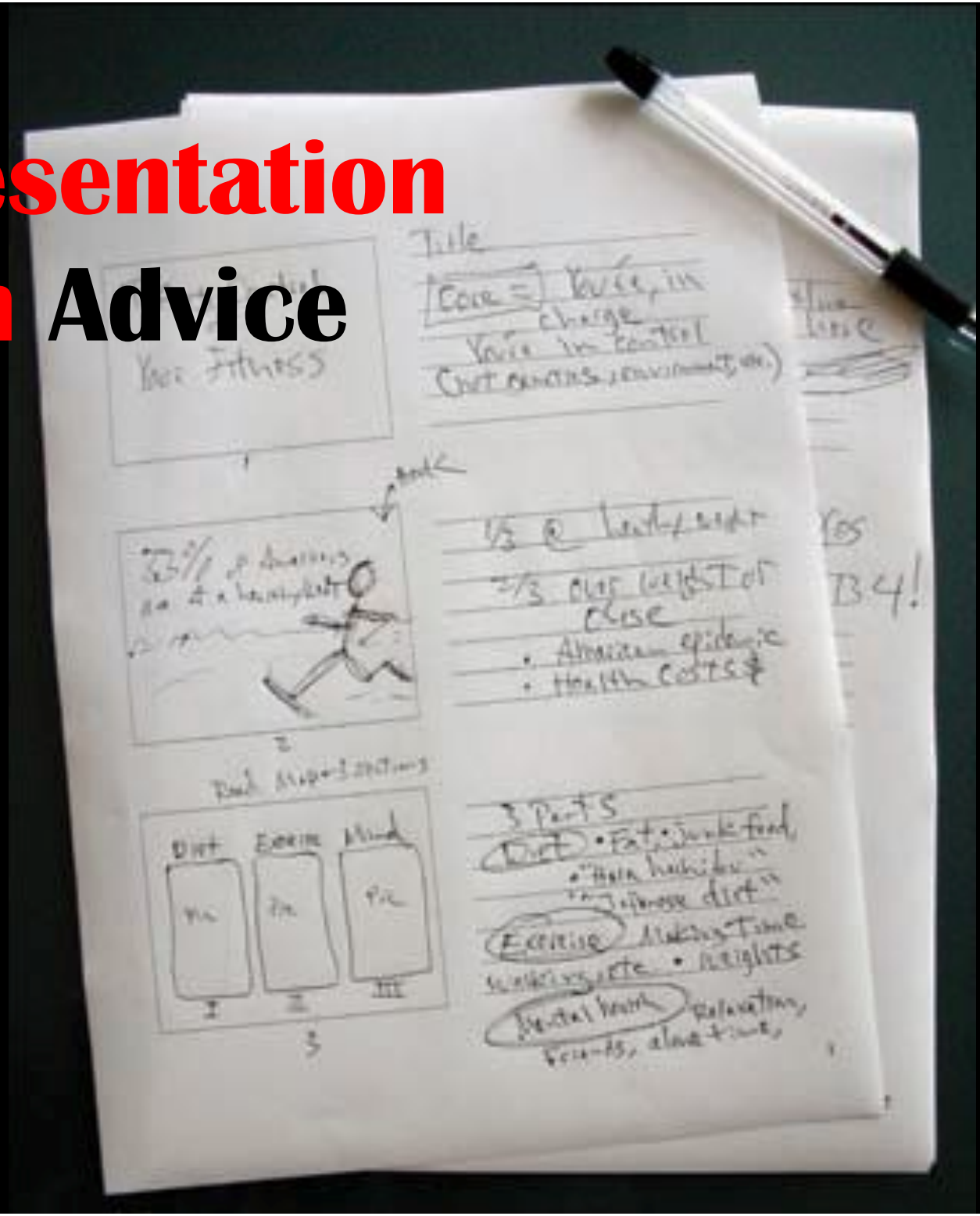


http://www.presentationzen.com

Presentation Zen Advice



Presentation Zen Advice



1

Best Practices for Linked Data Education

Tutorial at WWW 2013

Dr. Alexander Mikroyannidis
Knowledge Media Institute
The Open University, UK

2

The Open University

The Open University (OU) is the largest academic institution in the UK with:

- more than *250,000 students*
- close to *7,000 tutors*
- more than *1,200 full-time academic staff*
- more than *3,500 support and administrative staff*

Most OU courses are available throughout Europe and some are available worldwide.

Since its launch in 1969 more than *1.6 million people* worldwide have achieved their learning goals by studying with the OU.

3

Agenda

- The EUCLID project
- Best practices for curriculum design & delivery
- The EUCLID curriculum & module production process
- The EUCLID learning materials

4

Motivation

- There is a growing demand for Data Scientists possessing skills and detailed knowledge in Linked Data.
- Open Educational Resources (OERs) and MOOCs are changing the landscape of education online.
- The EUCLID project brings together these two developments by delivering a curriculum and learning materials with a focus on Linked Data.

5

The EUCLID project

- A European project facilitating *professional training for data practitioners*, who aim to use Linked Data in their daily work.
- EUCLID delivers a curriculum implemented as a combination of *living learning materials and activities* (eBook series, webinars, face-to-face training).
- The EUCLID curriculum is *validated by the user community* through continuous feedback.

6

EUCLID target groups

- *Practitioners* interested in using cutting-edge technologies to analyze and make sense of large amounts of Structured Data.
- *Technology enthusiasts* aiming to broaden their knowledge on one of most exciting and rapidly evolving topic related to Big Data and open-access policies.
- *Researchers in the Semantic Web community* who want to gain basic, down-to-earth knowledge about proven-and-tested technologies for using Linked Data.
- *Researchers in other disciplines* interested to use Linked Data technologies to optimize specific Data Management aspects related to the scientific data they operate on.

7

EUCLID partners

Core partners:

KIT | ontotext | research | The Open University

Associate partners:

Universidad Simón Bolívar | University of Southampton | fluid Operations

8

Best practices for curriculum design

- Industrial relevance
- Team curriculum design
- External collaboration
- Explicit learning goals
- Show realistic solutions
- Use real data
- Use real tools
- Show scalable solutions
- Eating our own dog food

9

Industrial relevance

- The EUCLID curriculum takes into account the needs of industry related to Open Data and Linked Data.
- Future work aims to automatically mine and analyse relevant job adverts to gain desired competencies for the sector.

10

Team curriculum design & external collaboration

- The EUCLID team is composed of a number of roles to fully capture *industrial*, *academic* and *pedagogical* requirements:
  - Industry (Ontotext, FluidOps) who have extensive experience with professional training, industrial requirements and scalable tools
  - Academia (KIT, STI International) who have research expertise in Linked Data and pedagogical experts (The Open University).
- External collaboration for gaining world-class curriculum expertise and for facilitating course delivery and dissemination.

11

Explicit learning goals

- All content (slides, webinars, eBooks) are developed based on a set of explicit learning goals.
- Learners are guided through the learning goals by learning pathways – a sequence of learning resources to achieve a learning goal.

12

Realistic solutions, real data & real tools

- Rather than toy examples we utilize systems which are deployed and used for real.
- We use a number of large datasets including for example, the MusicBrainz dataset which contains 100Ms of triples.
- Our collection of tools are used for real including for example: Seevl, Sesame, Open Refine and GateCloud.

13

Scalable solutions

- Based upon industrial-strength repositories and automatic translations.
- For example using the W3C standard R2RML for generating RDF from large data contained in standard databases.

14

Eating our own dog food

- We monitor communication and engagement with the Linked Data community through
  - W3C email lists
  - Social networking channels (LinkedIn & Twitter)
  - Content dissemination channels (Vimeo & SlideShare
- We transform the monitoring results into RDF and make these available at a SPARQL endpoint. In this respect we use Linked Data to support Learning Analytics.

15

Best practices for curriculum delivery

- *Open to format*
  Our learning materials are available in a variety of formats including: HTML and as an eBook, as an Apple iBook and on Amazon Kindles.
- *Addressibility*
  Every concept in our curriculum is URI-identified so that HTML and RDF(a) machine-readable content is available.
- *Integration*
  The main textual content, relevant webinar clips, screen casts and interactive components are placed into one coherent space.

16

Best practices for curriculum delivery

- *High quality*
  We have a formalised process where all materials go through several iterations to ensure quality, e.g. for each module we run both a practice and a full webinars facilitating critique and commentary.
- *Self-testing and reflection*
  In every module, we include inline quizzes formulated against learning goals enabling students to self-monitor their progress.

17

The EUCLID curriculum

A series of modules, each targeting a different crucial task related to Linked Data:

1. Introduction and Application Scenarios
2. Querying Linked Data
3. Providing Linked Data
4. Interaction with Linked Data
5. Creating Linked Data Applications
6. Scaling up

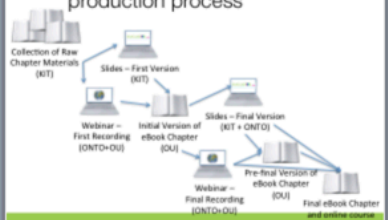
18

The EUCLID curriculum

- It has been designed to gradually build up trainee's knowledge.
- It enables trainees with previous knowledge on a specific area of interest to only briefly go over the introductory materials and directly dig into one of the more advanced modules.

19

The EUCLID module production process



20

The EUCLID learning materials

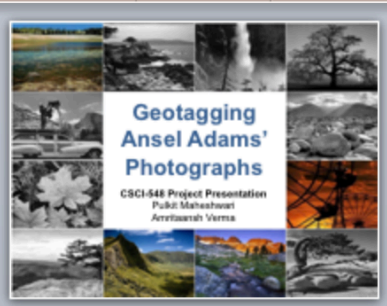
- Presentation slides
- Webinars
- Screencasts
- Exercises & quizzes
- eBook chapters
- Online courses

Geotagging.pptx

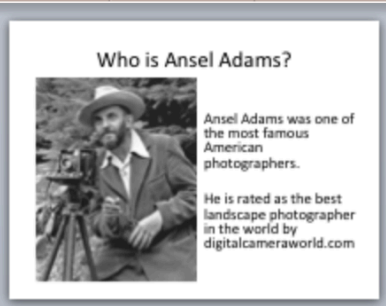
115%

Search in Presentation

Home | Themes | Tables | Charts | SmartArt | Transitions | Animations | Slide Show | Review



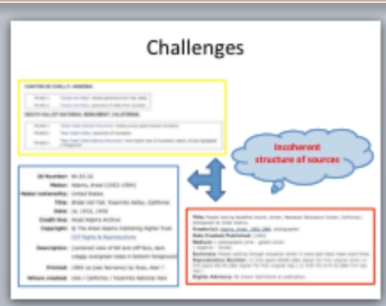
1 :20



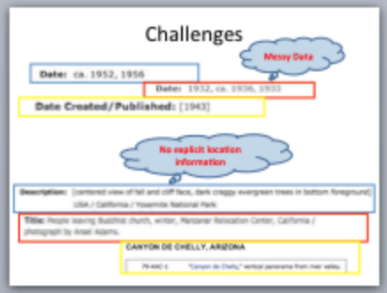
2 :20

3 :20

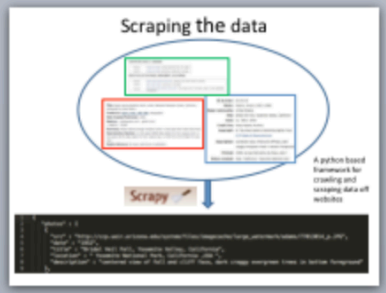
4 :20



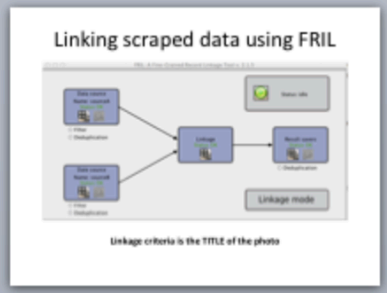
5 :20



6 :20

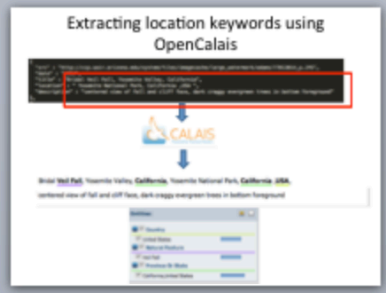


7 :20



8 :20

9 :20



10 :20

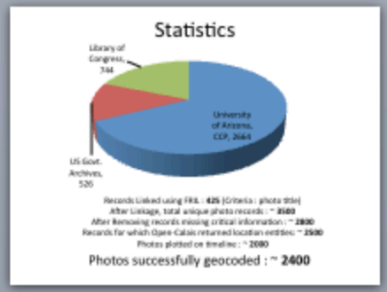
11 :20

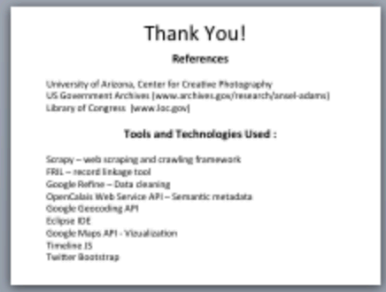
12 :20

13 :20

14 :20

15 :20





Slide Sorter View

115%

Project Video

4 minutes max

Narrated

Uploaded to YouTube

find your partner for the project

start thinking about your project

it takes time to come up with a good idea

http://pedroszekely.blogspot.com/2013/07/tips-for-working-successfully-in-group.html

Waiting List

- Even if you are in the waiting list, come to the classes and do the quizzes/homeworks
- Likely there will be additional openings in the next couple of weeks

When the Course is Over

- Directed research (MS or Phd Students)
- M.S. Thesis
- Summer interns (MS or Phd)
- Research Assistantships (1-2 Phd Students)
  - We can also recommend you for positions in other groups
- Graders/Course Producers (MS)
- Recommendation letters (anyone that gets at least an A-)
- Positions at related companies
  - Companies are often looking for recommendations of students

Questions?