UNIVERSITY OF Southampton
School of Electronics and Computer Science

Let Google Index Your Media Fragments

Yunjia Li, Dr Mike Wald and Dr Gary Wills

{yl2,mw,gbw} @ecs.soton.ac.uk

School of Electronics and Computer Science

University of Southampton

Agenda

UNIVERSITY OF Southampton
School of Electronics and Computer Science

- Media Fragments
- Problems of media fragments indexing
- The model to let Google index media fragments
- Discussion
- Conclusions

Find Charlie

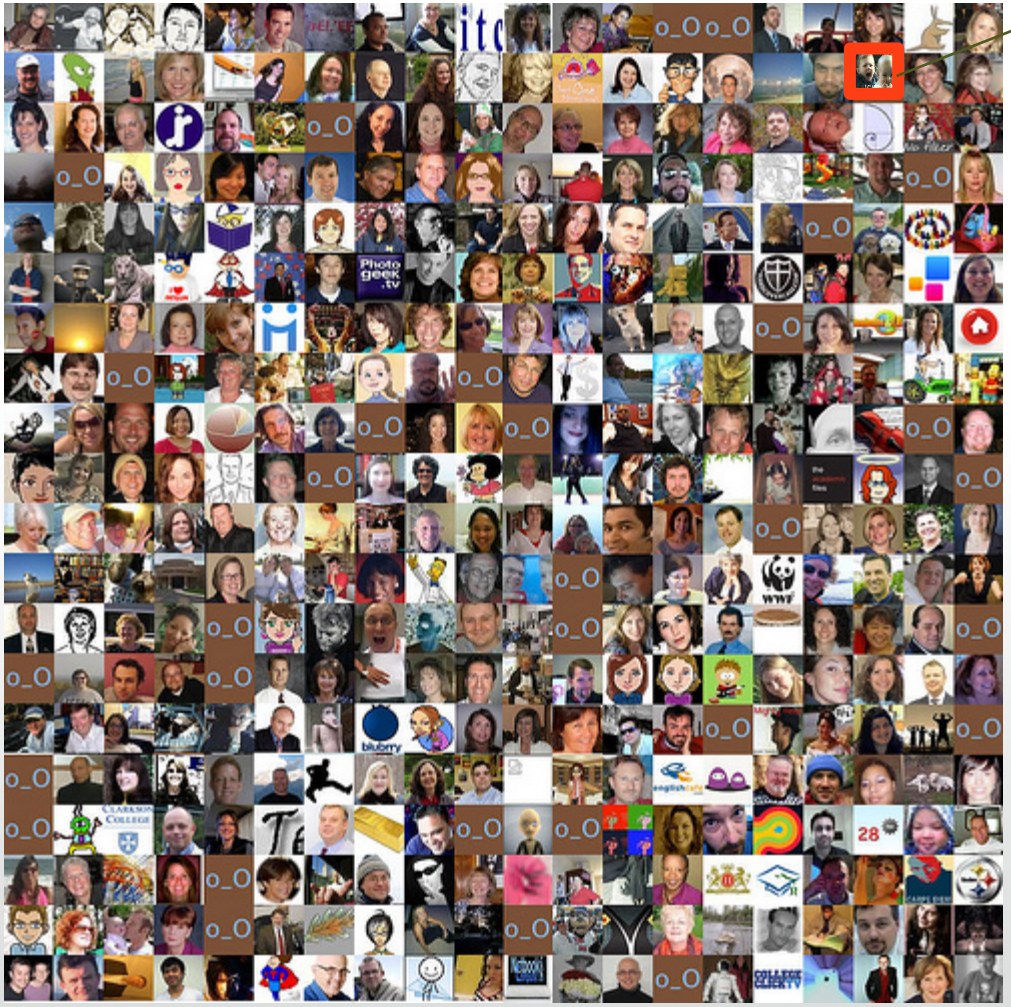
UNIVERSITY OF Southampton
School of Electronics and Computer Science

If you search “Charlie” in Google and find the following pictures picture in the result list…

Find Charlie (cont)

UNIVERSITY OF Southampton
School of Electronics and Computer Science

- Again, where is Charlie?



You may expect that the search engine can tell you where exactly is Charlie.

Media Fragment

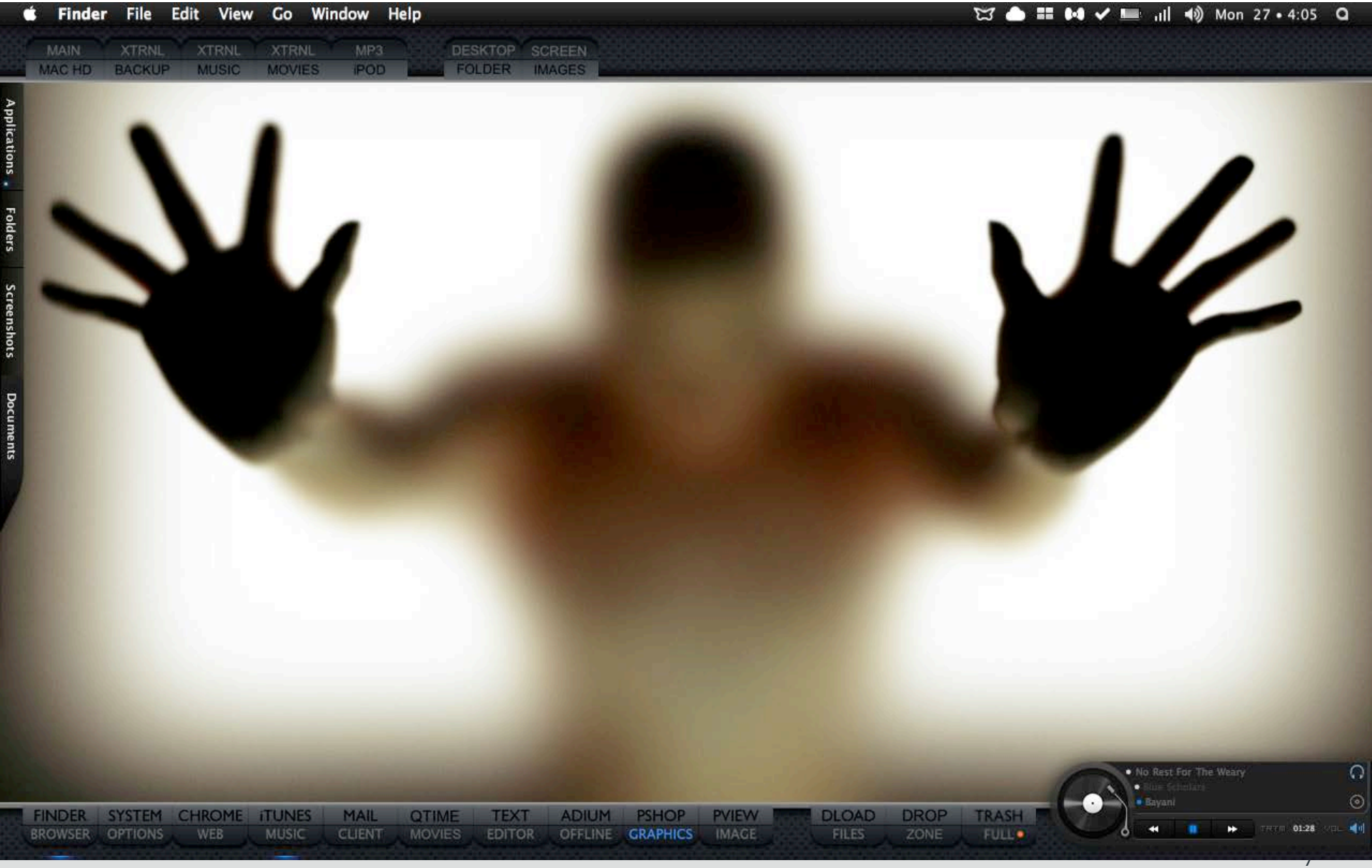
UNIVERSITY OF Southampton
School of Electronics and Computer Science

- Denote the inside content of multimedia resources
- Four dimension defined in Media Fragment 1.0 spec
  - Temporal dimension (10s to 20s of a 20mins long video clip)
  - Spatial dimension (a rectangle area in an image)
  - Track dimension (different sound tracks or captioning for different languages)
  - Named dimension (the combination of the above three)

Current Situation

UNIVERSITY OF Southampton
School of Electronics and Computer Science

- Multimedia uploading, sharing, tagging is easy
- Searching WHOLE multimedia resource is easy
- But searching PART of multimedia resource is difficult
  - Annotations are not linked to media fragments
  - Linked using javascript
  - no unique page for each media fragment
  - Example


Finder File Edit View Go Window Help
Mon 27 • 4:05
MAIN XTRNL XTRNL XTRNL MP3
MAC HD BACKUP MUSIC MOVIES iPOD
DESKTOP SCREEN
FOLDER IMAGES
Applications
Folders
Screenshots
Documents
FINDER SYSTEM CHROME iTUNES MAIL QTIME TEXT ADIUM PSHOP PVIEW DLOAD DROP TRASH
BROWSER OPTIONS WEB MUSIC CLIENT MOVIES EDITOR OFFLINE GRAPHICS IMAGE FILES ZONE FULL
No Rest For The Weary
Bayani
01:28

What we want to do...

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Title → The next Web of open, linked data

Presenter → Tim Berners Lee

06:02 → Linked Data Principles

07:28 → Grassroot picture

08:21 → DBpedia

09:15 → Government Data

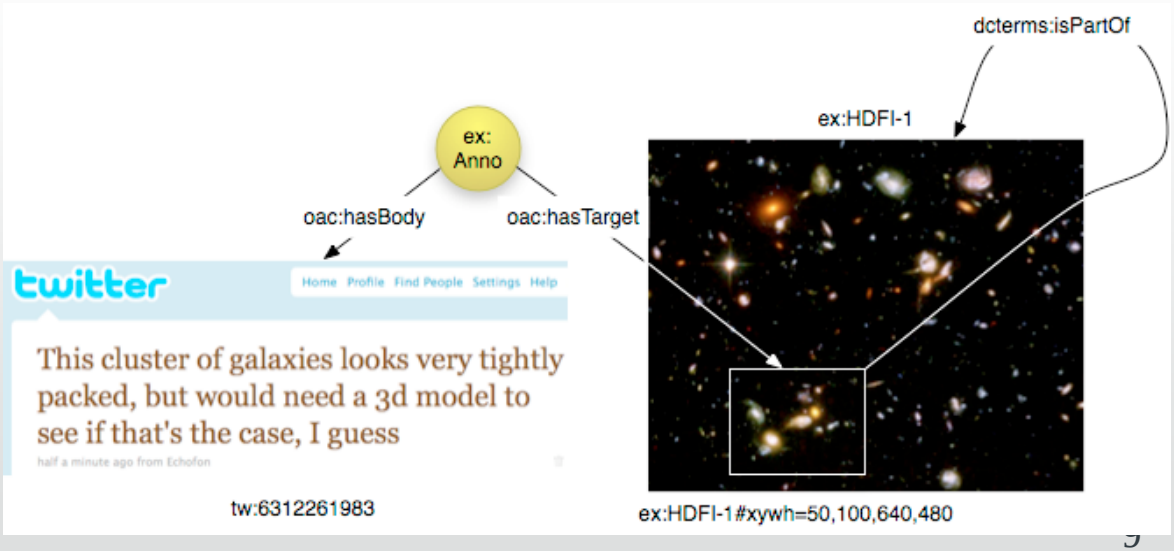
…

Gimme the FRAGMENT, not the WHOLE video!

The Difficulties

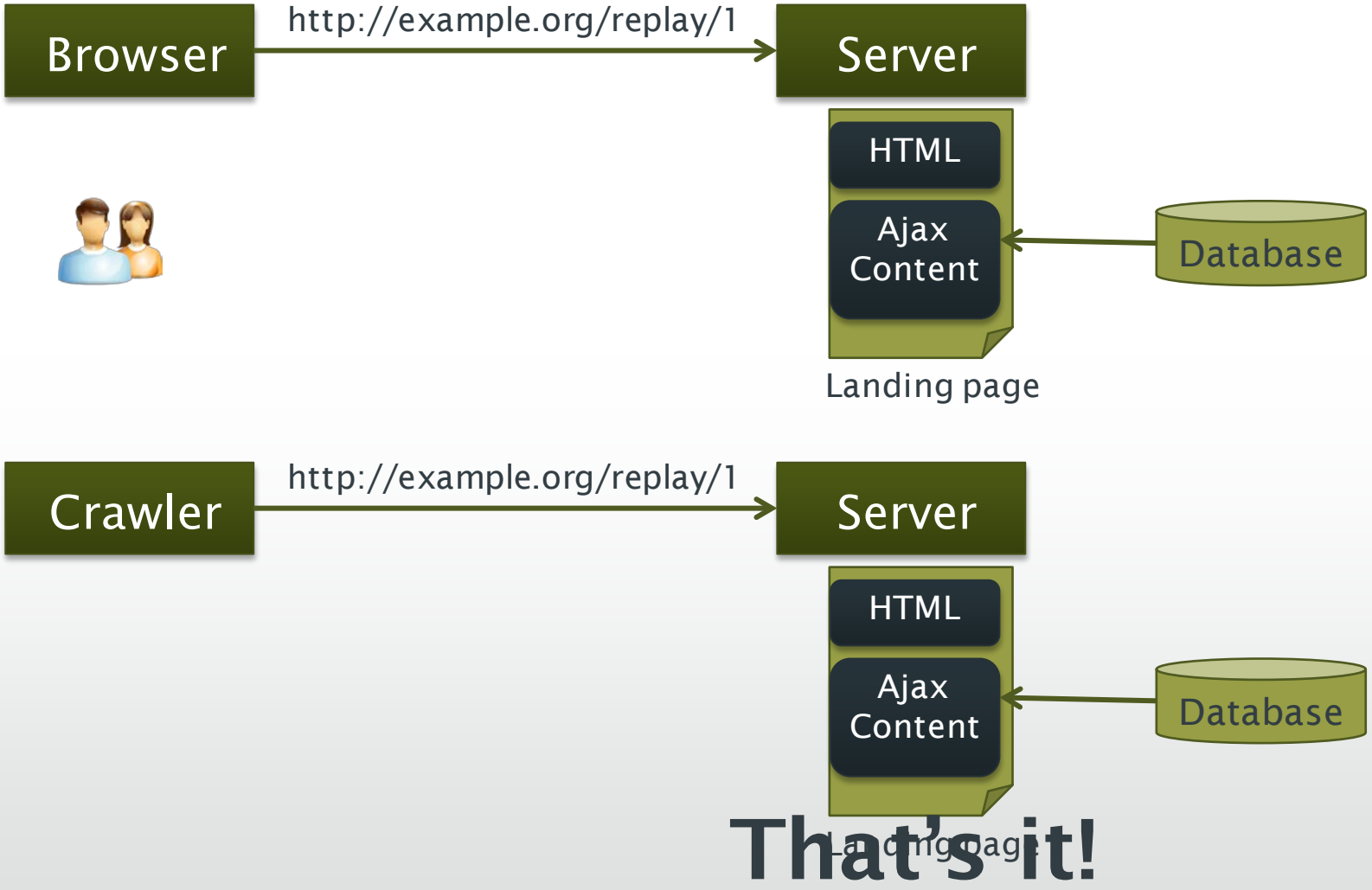
UNIVERSITY OF Southampton
School of Electronics and Computer Science

- The landing page is not search-engine-friendly
  - Everything is on the same page and the notion of media fragment is not explicitly embedded in HTML
  - Ajax content is ignored (e.g. Youtube interactive trans)
  - No semantic descriptions, no preview in search results



The Difficulties (cont)

UNIVERSITY OF Southampton
School of Electronics and Computer Science



That's it!

Google's Ajax Content Crawler

UNIVERSITY OF Southampton
School of Electronics and Computer Science

- The Crawler is designed to index Ajax content
- Replace token "#!" in URLs with "_escaped_fragment_"

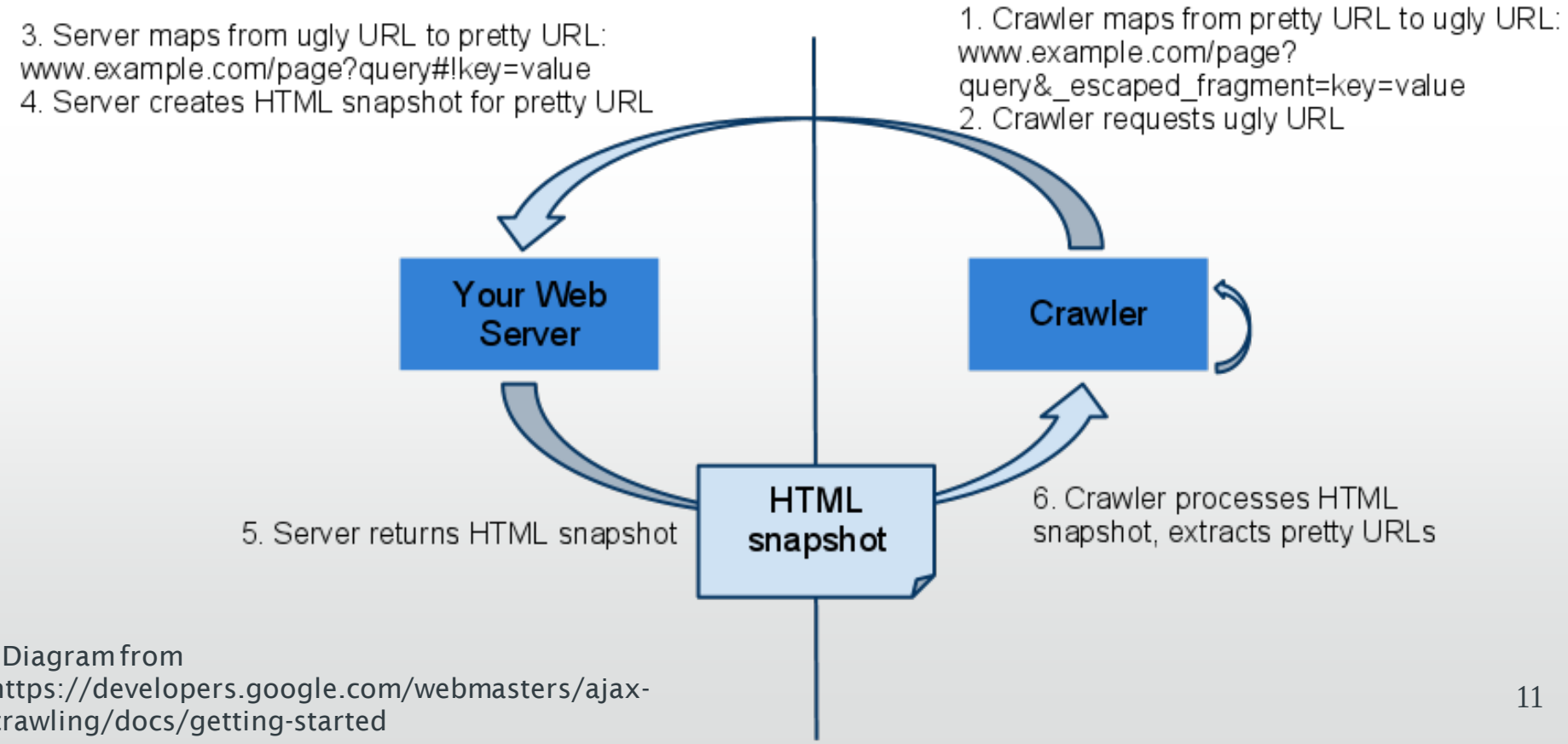


*Diagram from https://developers.google.com/webmasters/ajax-crawling/docs/getting-started

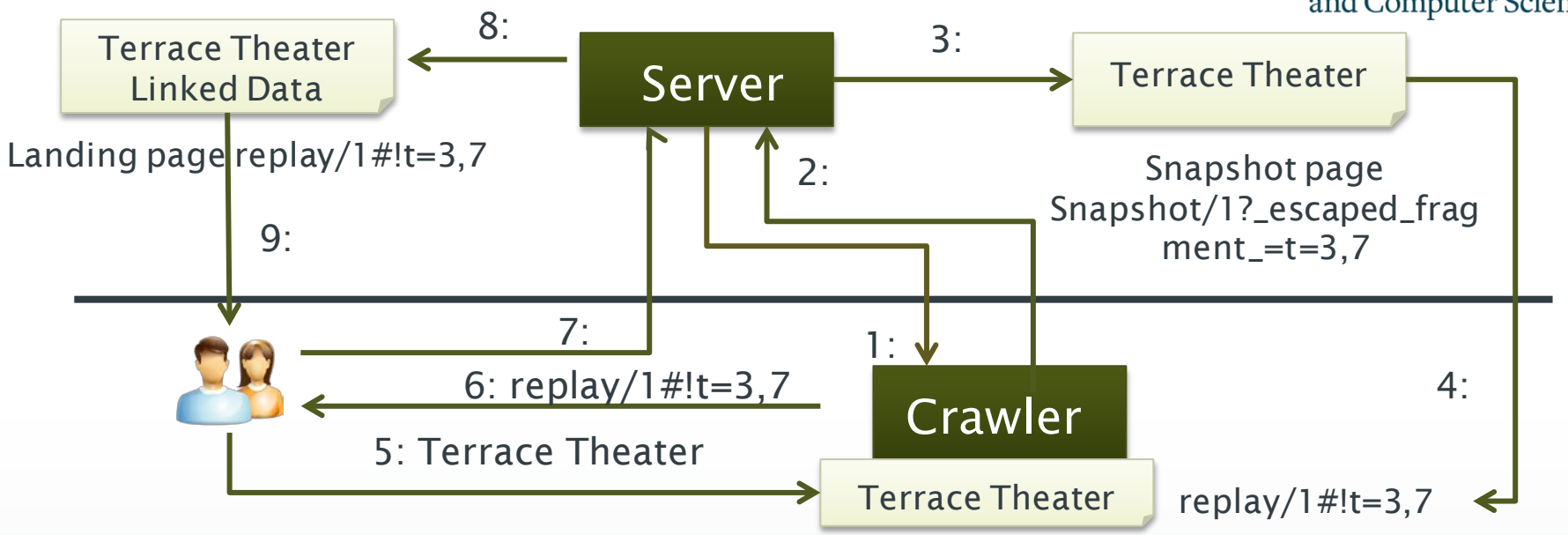
The Key Ideas in Solution

UNIVERSITY OF Southampton
School of Electronics and Computer Science

- The fragment information must be included in the URL
  - Syntax: W3C Media Fragment 1.0 Specification
- Prepare two sets of pages for every media fragment
  - Share the same landing page for users
  - the SEO page for search engine
- Landing page keeps the original user interaction
  - Highlight media fragments on opening
- SEO page
  - **ONLY** includes annotations of the media fragment
  - Embed rich snippet

The Solution

UNIVERSITY OF Southampton
School of Electronics and Computer Science



1: Submit pretty URL replay/1#!t=3,7 to the crawler

2: Crawler asks server for replay/1?_escaped_fragment_=t=3,7

3: Redirect the request to the snapshot page generated by the server. The snapshot page only contains annotations and Microdata for "#t=3,7",

4: The snapshot page is returned to the crawler with URL replay/1#!t=3,7

5: A user searches keyword "Terrace Theater"

6: Google includes replay/1#!t=3,7 in the search results

7: The user click the link and ask for the document at replay/1#!t=3,7

8: The server returns the landing page containing both "Terrace Theater" and "Linked Data"

9: The landing page highlights the media fragment by start playing from 3s to 7s

Introduction of Synote

UNIVERSITY OF Southampton
School of Electronics and Computer Science

- We implemented the model in Synote system
- User can generate annotations and synchronise them with audio-visual resources
- Synote doesn't store video, audio, image files
- Synote stores:
  - The URL references to video, audio image files online
  - User generated annotations and synchronisation points
- Single Resource: Tag, Note, Slide, etc
- Four categories of compound resources: Multimedia, Transcript, Synmark (tags, description), Presentation Slides

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Demo Time!

Synote v.s. TED Talks

All kinds of conceptual things, they have names now that start with HTTP

Discussion

UNIVERSITY OF Southampton
School of Electronics and Computer Science

- This is a hack! It now only works for Google
- The Media Fragment URI syntax is slightly modified
- Hashbang URL "#!" has many side effects [1]
- Microdata is embedded in snapshot pages for each media fragment
- Keep another set of snapshot pages for SEO and do not need to abandon the existing landing pages
- A better solution could be adding vocabularies about media fragment to schema.org

1. See http://dannythorpe.com/2011/02/09/side-effects-of-hash-bang-urls/

Workload

UNIVERSITY OF Southampton
School of Electronics and Computer Science

- For existing applications, you need
  - Includes hashbang in your url
  - URL mapping for “#!” and “_escaped_fragment”
  - Programmes to generate the snapshot page
  - Highlight the fragment when the page is opened

Conclusion

UNIVERSITY OF Southampton
School of Electronics and Computer Science

- Initial attempts to improve the online presence of media fragments
- Future work
  - For more search engines, Yahoo!, Bing
  - Consider other aspects of SEO
- More media fragments could be published to benefit search

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Questions?