Organisation des Nations Unies pour l'éducation, la science et la culture

Chaire UNESCO en technologies pour la formation des enseignants par les ressources éducatives libres, Université de Nantes, France

Modèles de ressources éducatives libres : données, traitements et problèmes

Colin de la Higuera

AILE

13 mai 2019

Plan

1. Les enjeux
2. Le projet X5-GON
3. Les données brutes
4. Les données stockées
5. Les données enrichies
6. Les modèles de contenu
    1. La qualité
    2. La difficulté
    3. L’accessibilité
    4. La connexion temporelle

1. Les enjeux

1.1 L'éducation, des enjeux à l'échelle de la planète

- En 2012, dans certains pays d'Afrique de l'Ouest on pouvait compter 12 enfants pour un livre
- Il manque 4 universités de 30 000 étudiants par semaine d'ici 2025
- Au Brésil, 400 Millions de dollars d'argent public dépensé en livres scolaires par an (K-12)
- Les grands défis pour l'éducation de l'Unesco pour 2000 ne sont pas atteints en 2017
- Au Kenya, en Tanzanie ou en Ouganda, 75% des élèves de CE2 ne sait pas lire "The name of the dog is Puppy". En Inde Rurale 50% des élèves de CM2 ne peut pas faire une soustraction à deux chiffres comme 44-15.
- Au rythme actuel il faudra au Brésil 75 ans pour atteindre le niveau en mathématiques des élèves des pays riches, 260 ans pour la lecture.

- https://openknowledge.worldbank.org/bitstream/handle/10986/28340/211096ov.pdf
- http://www.fnde.gov.br/programas/programas-do-livro/livro-didatico/dados-estatisticos

Les challenges et opportunités

- 4.5 milliards d'abonnements aux téléphones portables dont 700 millions en Afrique
- 1.5 milliards d'abonnements incluent l'internet
- 2.4 milliards de Smartphones
- Les modes d'apprentissage qui se développent : *informal learning*, *mobile learning*, *offline learning*

Les acteurs globaux

- L'éducation se **globalise**
- L'*agence* qui aujourd'hui distribue le plus d'accréditations : coursera
- Les grandes entreprises du numérique
- La communauté européenne
- L'Unesco
- Les ONGs

Plus proche de nous, la réforme du Lycée (février 2019)

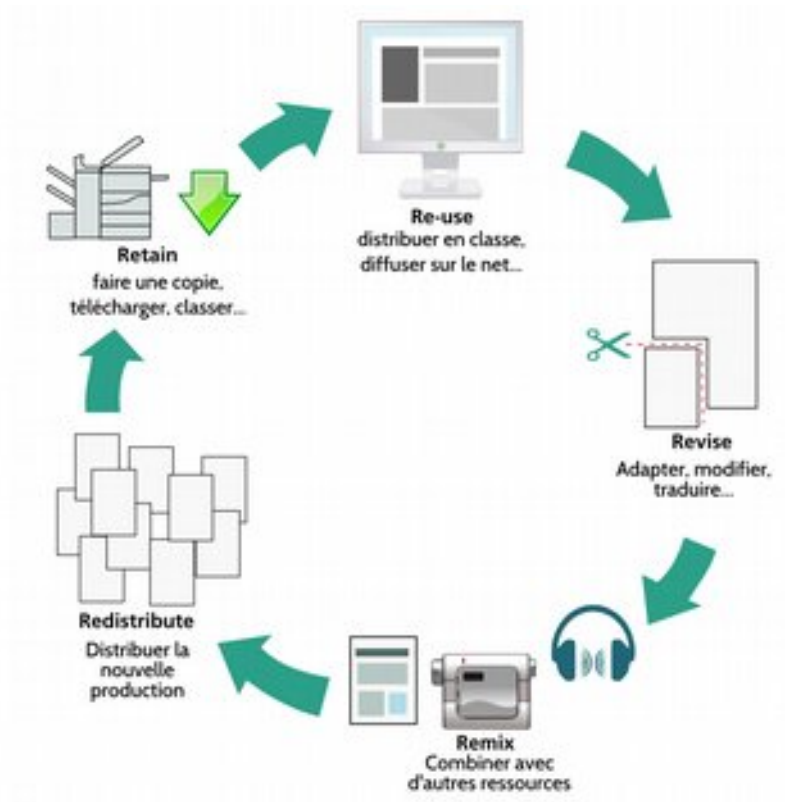
- À l'été 2018, François Bonneau, président délégué des Régions de France, avait estimé le coût du remplacement des manuels à 300 millions d'euros sur 2 ans, s'il se faisait uniquement au format papier.
- [...]
- Le manuel numérique relève en effet d'un modèle économie différent du papier, puisque les régions font l'achat d'une licence, à payer sur plusieurs années, ce qui permet de lisser le coût du manuel sur 4 ou 5 ans, selon les précisions de Célia Rosentraub.

"Le passage au numérique peut être une contrainte pour certains territoires ruraux"

1.2 Les ressources éducatives libres

Les 5 R

http://www.opencontent.org/definition/

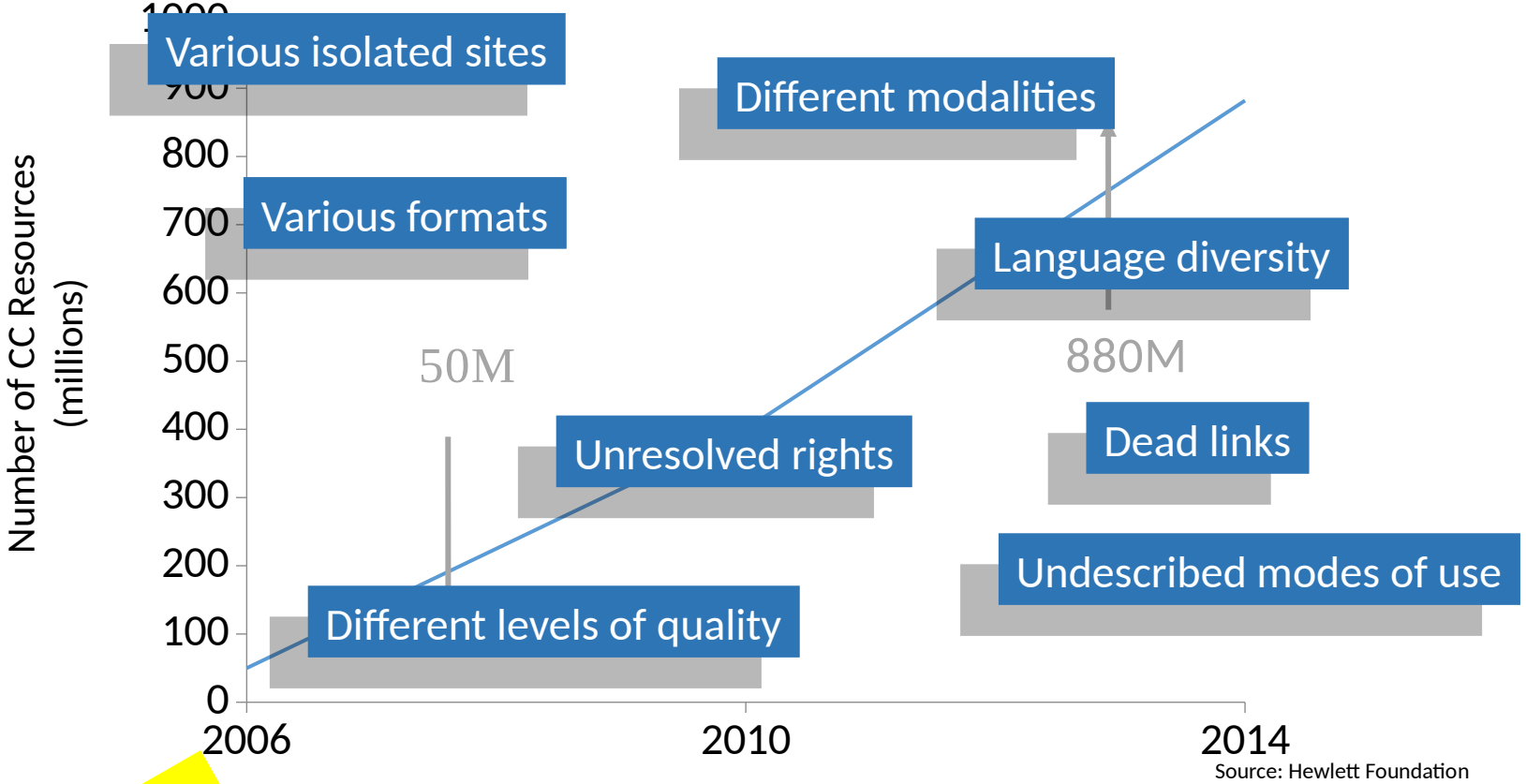


Les 5 R

- Retenir
- Réutiliser
- Réviser
- Remixer
- Redistribuer

2. Le projet X5-GON

Mass content missing mass use



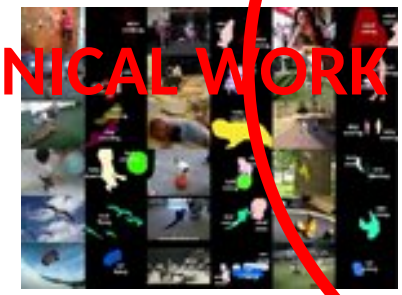
Mitja Jermol 2018

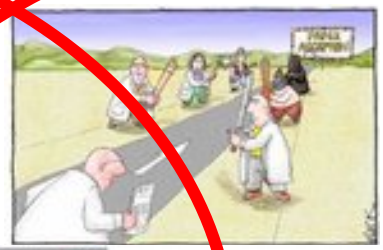
The Goal

X5gon is an **analytic platform** with **open services, APIs and scripts** supported with **AI** enabled technical pipeline to converge distributed OER content and media with users into a **one-stop-shop data-driven learning environment**.

- Cross-site: providing technologies to transparently accompany and analyse users across sites;
- Cross-domain: providing technologies for cross domain content analytics;
- Cross-modal: providing technologies for multimodal content understanding;
- Cross-language: providing technologies for cross lingual content recommendation;
- Cross-cultural: providing technologies for cross cultural learning personalisation.

Learning resources open pipeline





TECHNICAL WORK

RESEARCH WORK

Automatic ingestion, cleaning, fusion, preprocessing

Cross-lingual, cross-modal Semantic processing

Resources quality processing

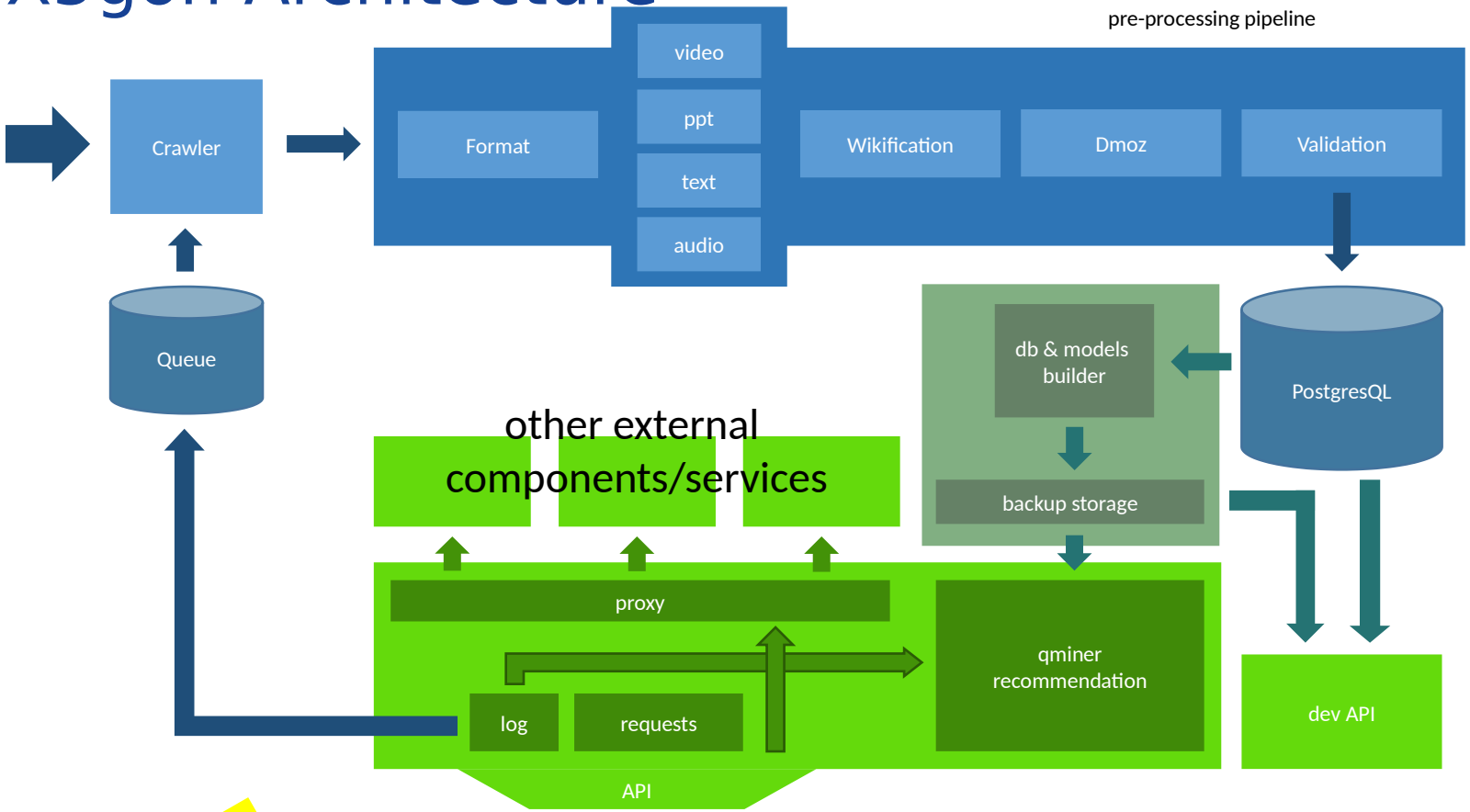
Didactic design and pedagogical processing

Innovation

Research

Mitja Jermol 2018

X5gon Architecture



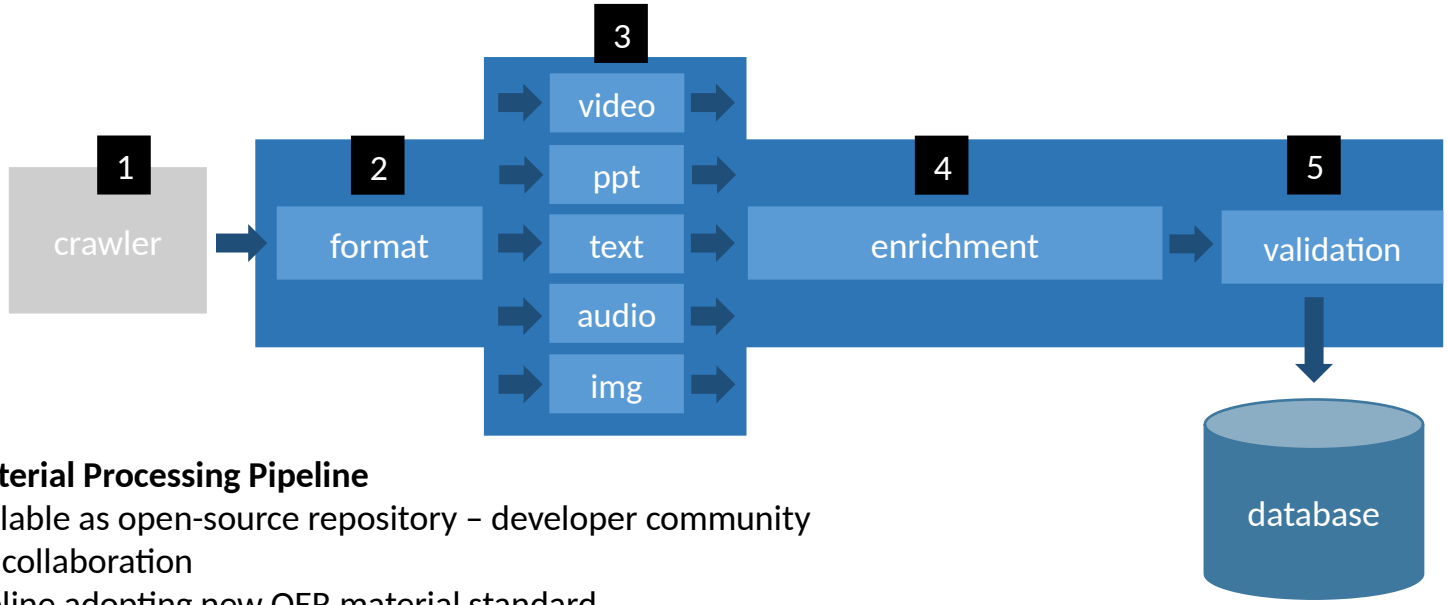
Mitja Jermol 2018

Who are we?

Participant No *	Participant organisation name	Country
1 (Coordinator)	UNIVERSITY COLLEGE LONDON	UK
2	INSTITUT JOZEF STEFAN	SI
3	KNOWLEDGE 4 ALL FOUNDATION LBG	UK
4	UNIVERSITAT POLITECNICA DE VALENCIA	ES
5	UNIVERSITÉ DE NANTES	FR
6	UNIVERSITAET OSNABRUECK	DE
7	POSTA SLOVENIJE	SI
8	MINISTRY OF EDUCATION OF SLOVENIA	SI

Mitja Jermol
2018

Open-Source Community



OER Material Processing Pipeline

- Available as open-source repository – developer community and collaboration
- Pipeline adopting new OER material standard
- Easy incorporation and usage within existing OER platforms
- Information validation – check if the required information is present
- Full documentation of the platform

Source code: github/x5gon
Documentation: x5.github.io

3. Les données brutes

Deux grands types de données

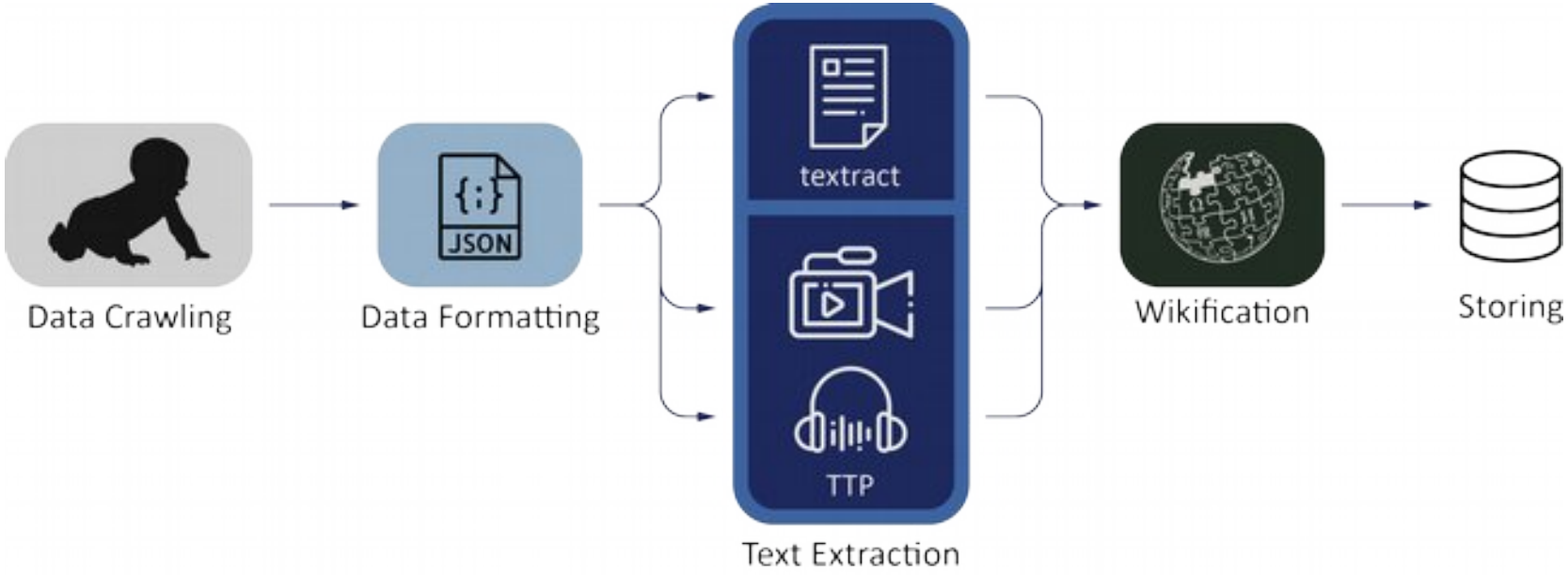
- Les données ressources :
  - Une ressource est une vidéo, un pdf, un audio, un texte
  - Une ressource pourrait être un quiz, un jeu, une carte, une image
- Les données usage
  - L'utilisateur u accède à la ressource R au temps t
  - L'utilisateur u consomme la ressource R d'une façon spécifique
- Un grand absent : les données utilisateur

Comment ces données sont obtenues

- Les données ressources sont obtenues par un crawler
- Les données utilisateur sont obtenues grâce à l’aide d’un snippet

Processing Pipeline
Overview

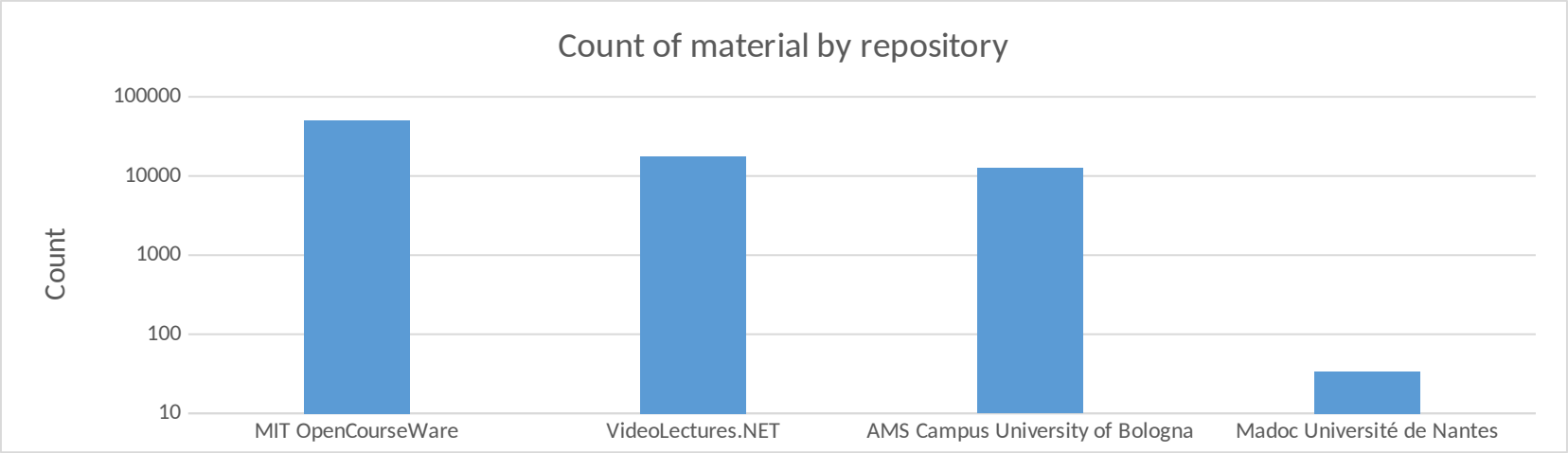
Based on the initial architecture [1]



[1] Deliverable 2.1: Requirements & Architecture Report

Material Data
Providers Statistics

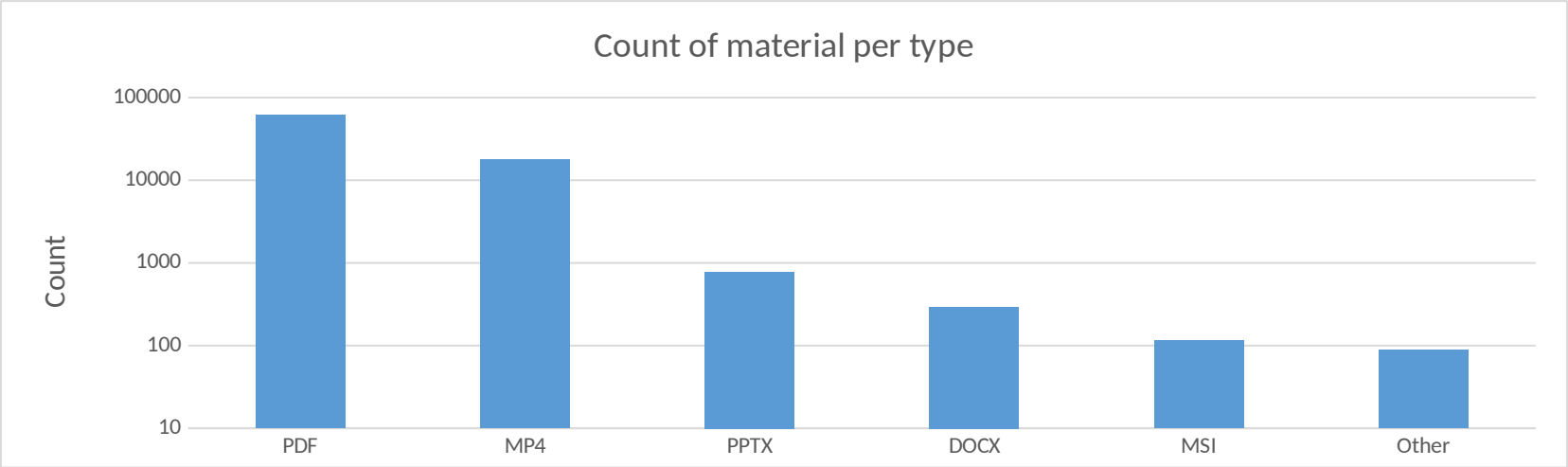
- Acquired and pre-processed approx. 90k items



- Most materials were acquired from MIT OpenCourseWare
  - MIT has mostly text materials – book chapters, problems, exams
  - MIT's resources are mostly in English

Material Data
File Types Statistics

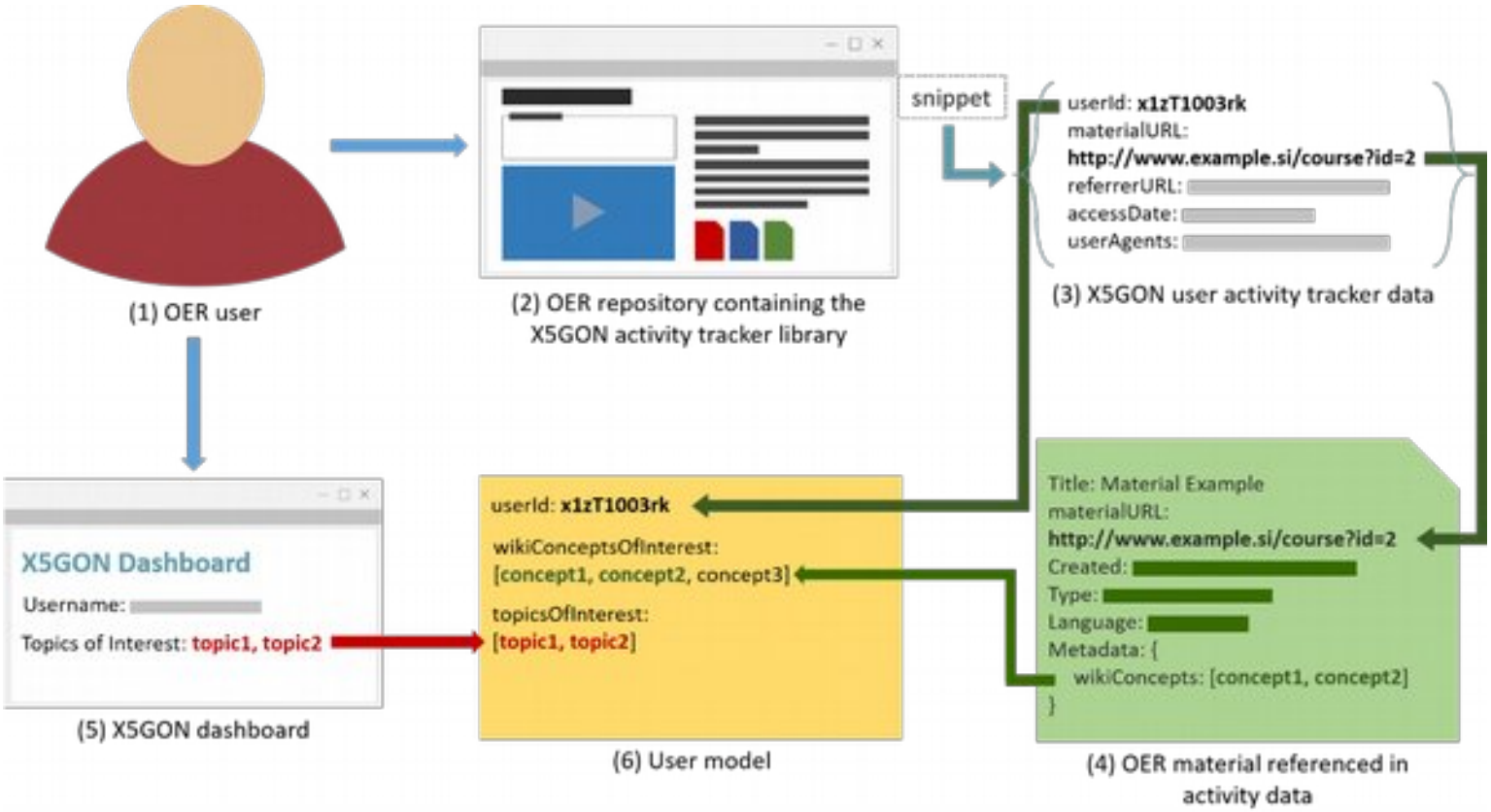
- Each file type can be represented in various formats



- Most dominant type – text (pdf, pptx, and docx)
  - Generalizing: Most of the OER are in some text format

User Activity Acquisition

Workflow



User Activity Acquisition
Integration

Currently integrated

- VideoLectures.NET
- Universitat Politecnica de Valencia
- Université de Nantes
- Universitaet Osnabrueck – just recently

Integration in process

- University College London

Integration in the future

- Other 3rd party repositories

User Activity Acquisition
Statistics

Repository distribution

Provider	All Visits	Known Visits	Unknown Visits	Unknown Visits : All Ratio	Unique Users
All	675,685	307,004	368,681	**0.5456**	141,544
Videolectures.NET	578,775	305,573	273,202	0.4720	141,429
PoliMedia	93,379	387	92,992	0.9959	16
Nantes	394	380	14	0.0355	31

Unknown visits ratio is high

- Identify crawling bots
- Detect technology that blocks 3rd party cookies

User Activity Acquisition

User Agents Statistics - Bots

Unknown user agents

- Highlight user agents that contain word “bot” or “preview”
- More than half of unknown visit logs are from bots

User Activity Acquisition

User Agents Statistics

Browsers

- What browser the users use to access material
- Safari disable 3rd party cookies by default

User Activity Acquisition

User Agents Statistics

Operating System

- What operating system the users use
- Mac OS and iOS default browser is Safari

User Activity Acquisition

User Pathway Graph

Where the users come from?

- Only known users
- Websites from which the user came to the OER – referrer

List of top referrers

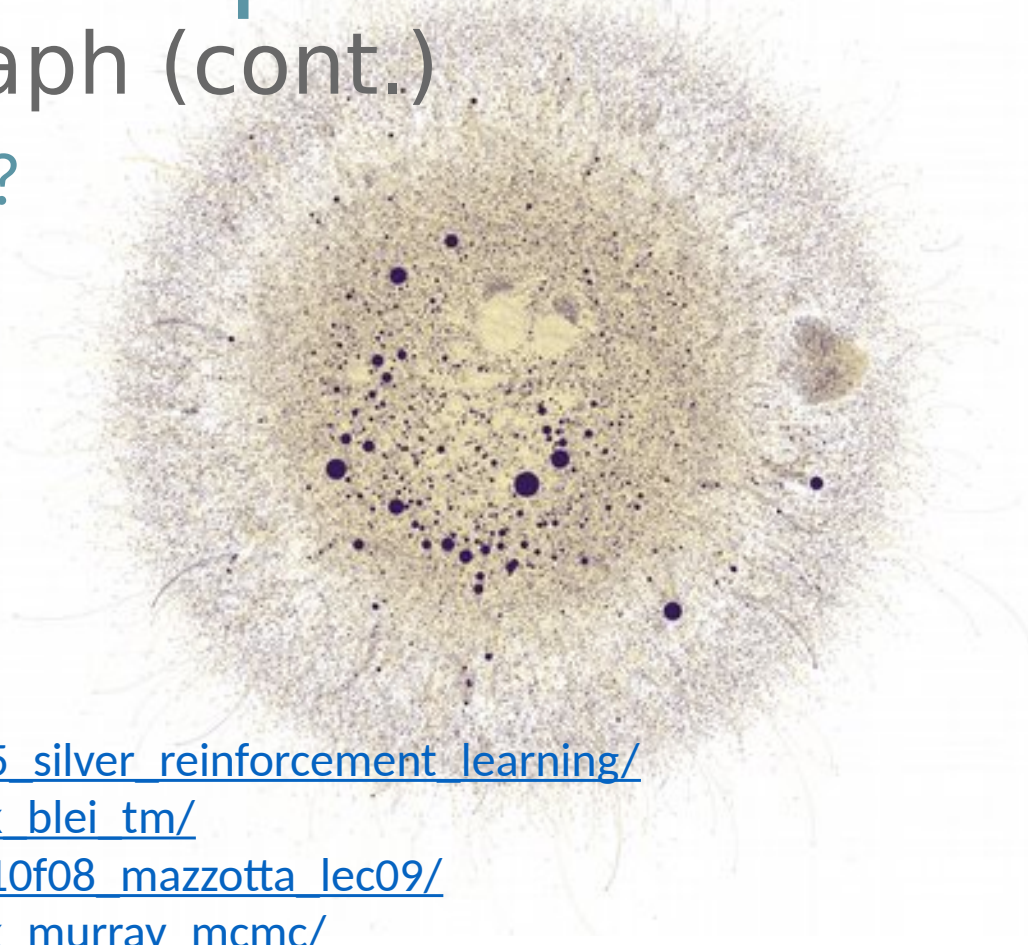
- http://google.com/
- http://videolectures.net/deeplearning2017_montreal/
- https://media.upv.es/
- http://videolectures.net/mit1806s05_linear_algebra/
- http://videolectures.net/mit801f99_physics_classical_mechanics/

User Activity Acquisition

User Pathway Graph (cont.)

Where the users end up?

- Only known users
- Lectures frequently visited by the users - materialURL



List of top materialURL

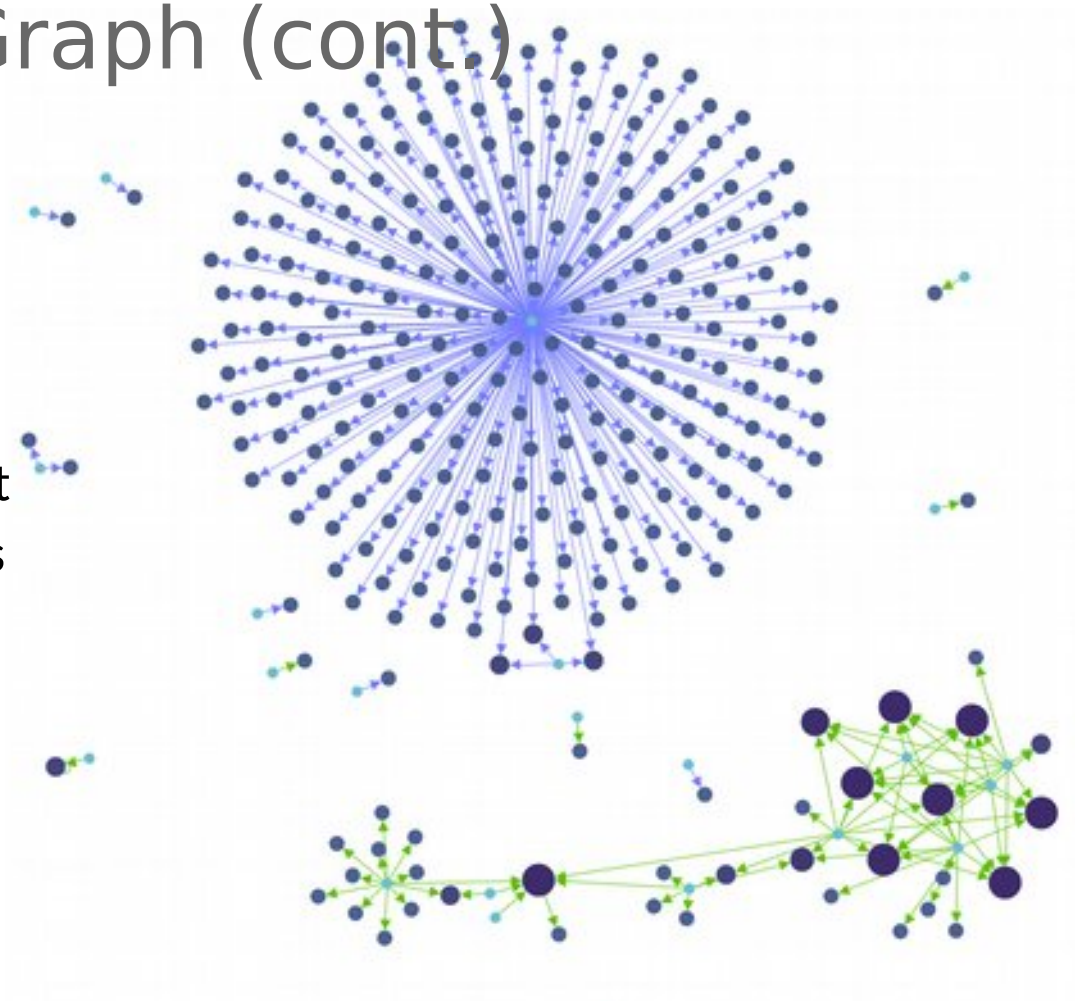
- http://videolectures.net/rldm2015_silver_reinforcement_learning/
- http://videolectures.net/mlss09uk_blei_tm/
- http://videolectures.net/yaleital310f08_mazzotta_lec09/
- http://videolectures.net/mlss09uk_murray_mcmc/
- http://videolectures.net/deeplearning2017_precup_machine_learning/

User Activity Acquisition

User Pathway Graph (cont.)

Activity of particular users

- Two users and their activities
  - Arrows show movement
  - Edge color distinct users



Les enjeux RGPD

- Pas d'enjeu sur les données ressource. L'enjeu est celui des licences.
- Une licence CC-BY, CC-BYSA, CC-BYNC permet de partager les ressources.
- Par défaut, une ressource n'est pas partageable
- Remarque : ça peut ne pas être un frein à sa recommandation !

- Gros enjeu sur les données usage. Le site qui « offre » ses données doit prévenir (par un bandeau) et offrir toutes les informations
- Beaucoup de questions
  - Prévenir à chaque fois ?
  - Suivre par inscription ou par cookie ?

4. Les données stockées

X5gon-TTP (cont.)

Total OER videos transcribed and translated per site and language

	VL.NET	poliMedia	VirtUOS
English	1629	165	26
Spanish		3870	
German		4	250
Slovene	391		

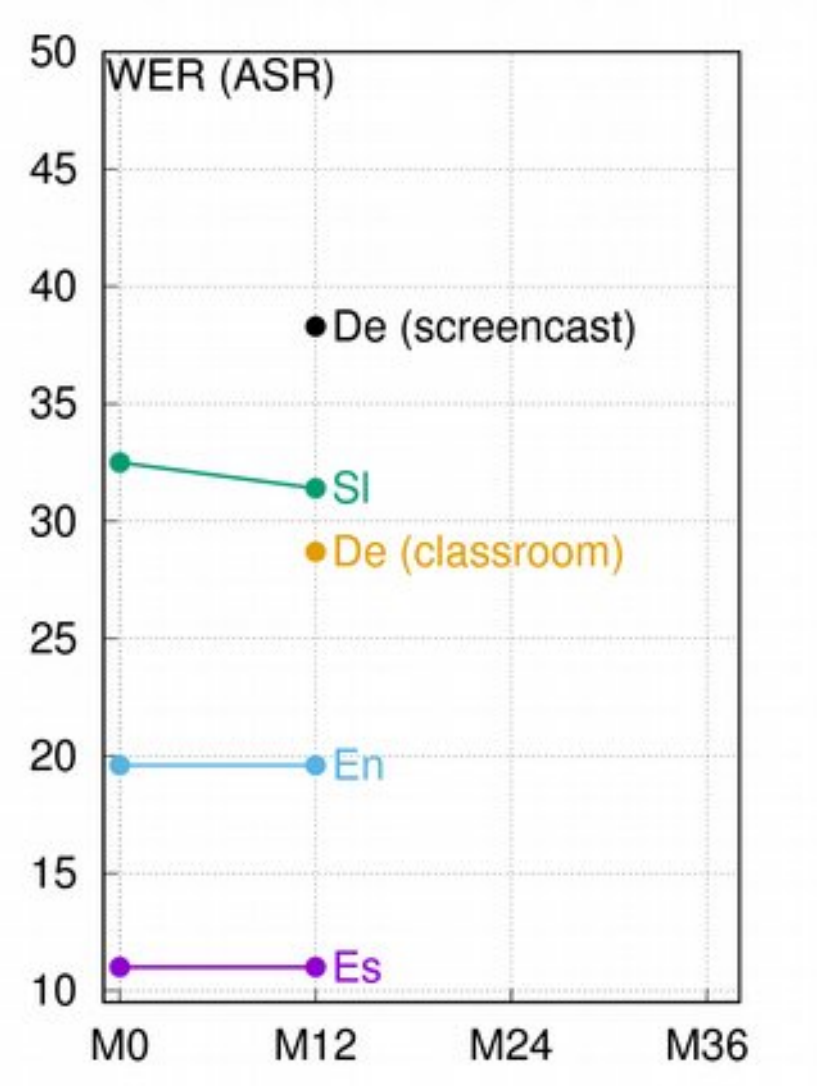
X5gon-TTP (cont.)

· Subtitle Editor for reviewing automatic transcriptions and translations

First results on transcription quality

ASR test sets:

Lang.	Set name	Videos	Hours
es	poliMedia	23	3.0
en	VL.NET	4	3.4
sl	VL.NET	4	3.4
de	VirtUOS (scr. cast)	19	5.3
de	VirtUOS (class)	2	2.4



5. Les données enrichies

Understanding content

...is...

- transcription and translation
- metadata extraction and collection
- topic modelling of the resources
- building models for the OER
- external topology of these resources (similarities and distances)
- internal topology of these resources (understanding the internal organization)
- computation of indicators (quality, complexity)

Travail Mica Ménard / Anthony Allaire 2018, Victor Connes 2019

Mica & Anthony, work 1

- Given a video, compute the vector
- Compute the distance between 2 vectors

- Plusieurs alternatives
- Bag of words / TF-IDF
- Embeddings
- Wikifier/wikification

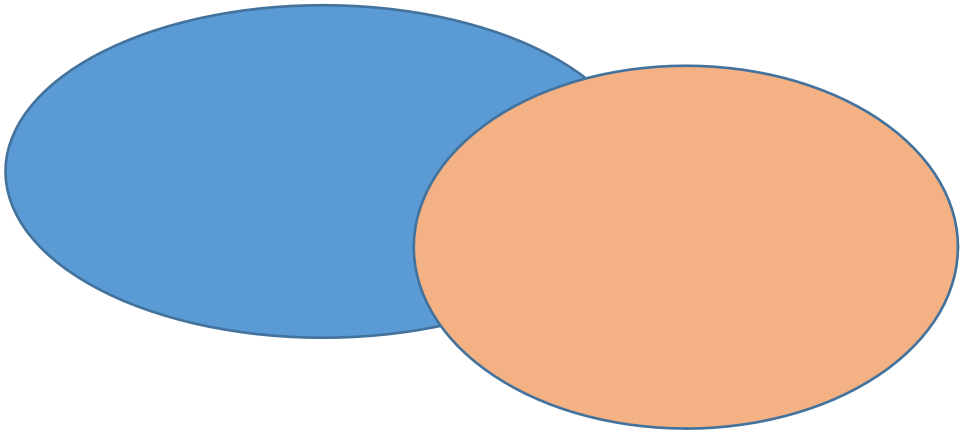
Travail Anthony Allaire 2018, Victor Connes 2019

Find the missing lecture!

- Goal: to be able to find the missing video.
- Task: we are given two resources RA and RB. Find an intermediate resource RX from a repository which somehow bridges the learning gap from RA to RB.
- Obviously the problem is ill defined. So there is an issue about actually understanding the problem
- For a first task we imagine an identification task. RA and RB are two lectures from a series of lectures (for example lecture 4 and lecture 6). The repository contains the other lectures from the collection and many other resources. The goal is therefore to identify lecture 5.

First results

- The technique consists in computing the TF-IDF score of $(RA \cup RB) \cap RX$. This represents the elements which are in common



Corpora	C1	C2	C3	C4
Top choice	60 %	38,46 %	40 %	57,14 %
Top 10	90 %	100 %	95 %	100 %

C1 : Introduction to Computer Science and Programming in Python: https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-0001-introduction-to-computer-science-and-programming-in-python-fall-2016/lecture-videos /

C2 : Introduction to Computational Thinking and Data Science: https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-0002-introduction-to-computational-thinking-and-data-science-fall-2016/lecture-videos /

C3 : Foundations of Computational and Systems Biology: https://ocw.mit.edu/courses/biology/7-91j-foundations-of-computational-and-systems-biology-spring-2014/video-lectures /

C4 : Freshman Organic Chemistry I: https://oyc.yale.edu/chemistry/chem-125a

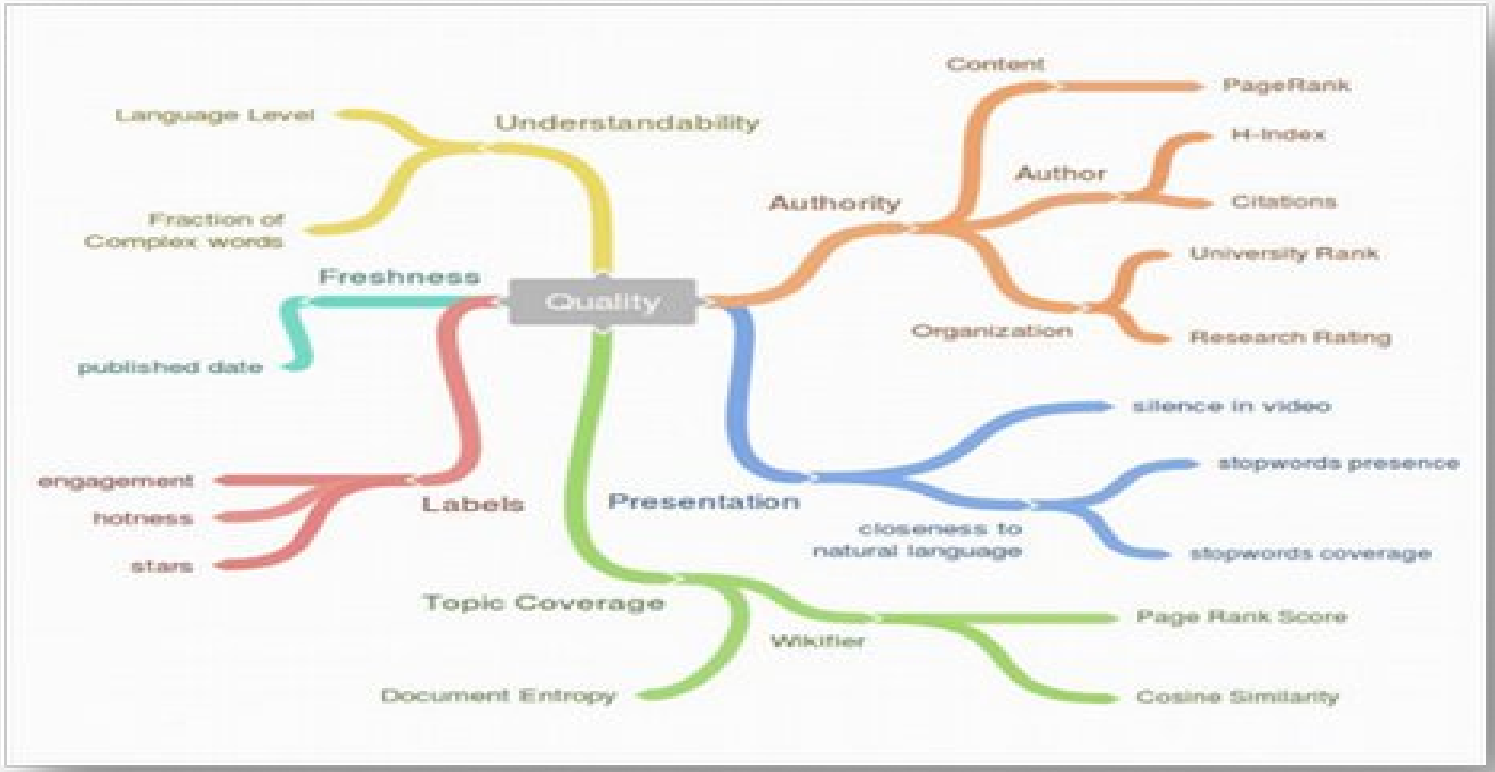
6. Les modèles de contenu

6.1 La qualité

Travail effectué à Londres

Sahan Bulathwela (UCL)

Different Dimensions of Quality



Explicit vs. Implicit Labels of Quality

- 3, 014 lectures with at least a single star rating
  - Not sure how many users rated a lecture
  - 0 stars to 5 stars

- 25, 230 lectures with hotness score

$$Hotness = \frac{Number\ of\ views}{Number\ of\ days\ since\ publication^2}$$

- 14, 877 lectures with at least one engagement data point

$$Engagement\ Rate = \frac{Total\ Duration\ of\ lectures\ watched\ by\ learner}{Length\ of\ lecture}$$

[Meyerson (2012)]

  - 6, 270 lectures with 5 or more user views
  - 3, 223 lectures with 10 or more user views

6.2 La difficulté

Les challenges

- Mesurer la difficulté relative entre deux ressources
- Savoir si la ressources A peut / doit être « consommée » avant la ressource B

- Mesurer une difficulté absolue d'une ressource.
- Méthode ad hoc 2018

1. Construire le vecteur TF-IDF associé à une ressource
2. Trier ce vecteur. Normaliser
3. Calculer la kurtosis
4. Celle-ci est un indicateur de complexité

6.3 L’accessibilité

Problème ouvert

Comment définir (automatiquement) si une ressource est accessible à tous ?

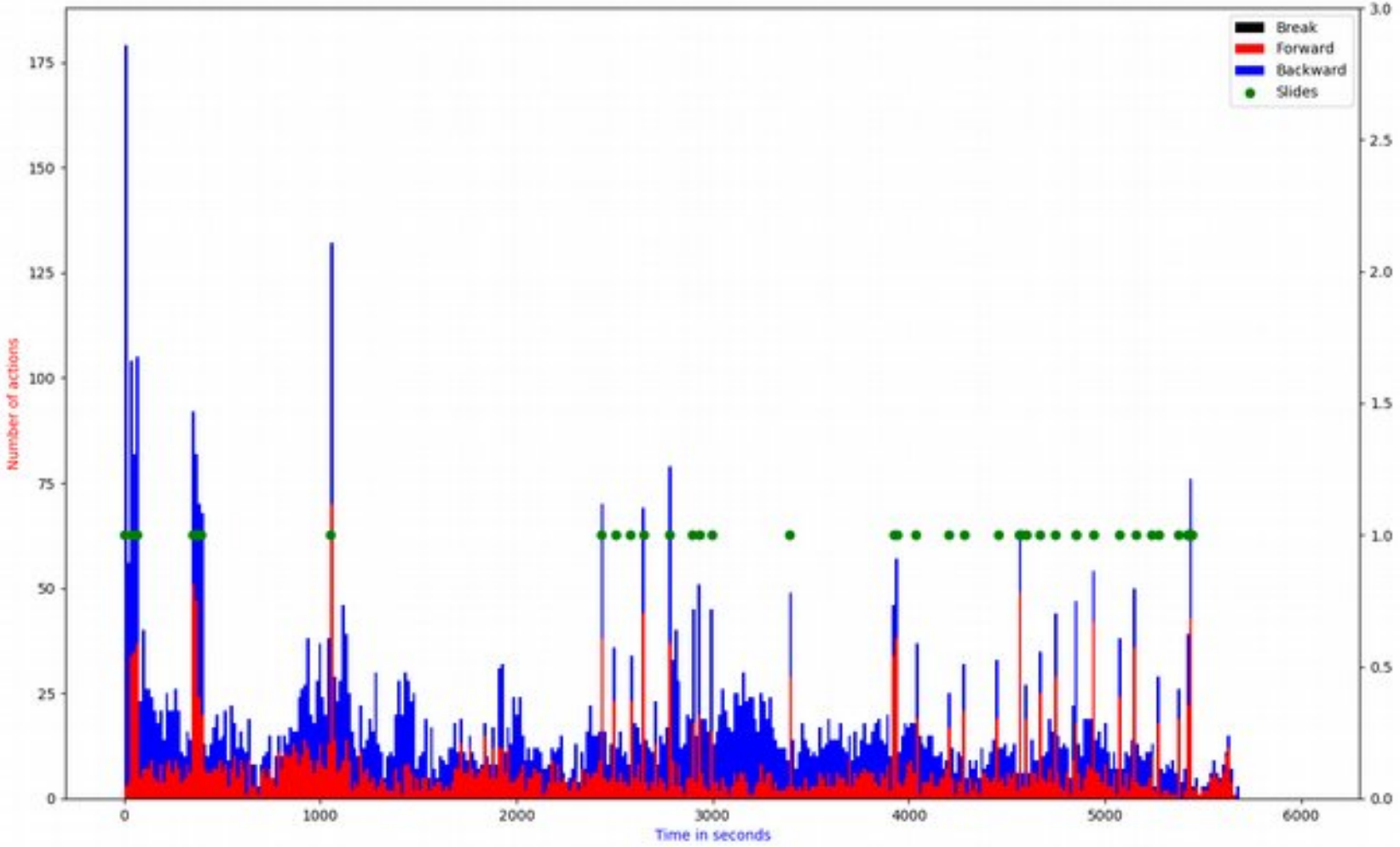
Peut-être est-il plus simple de mesurer si une ressource pose problème ?

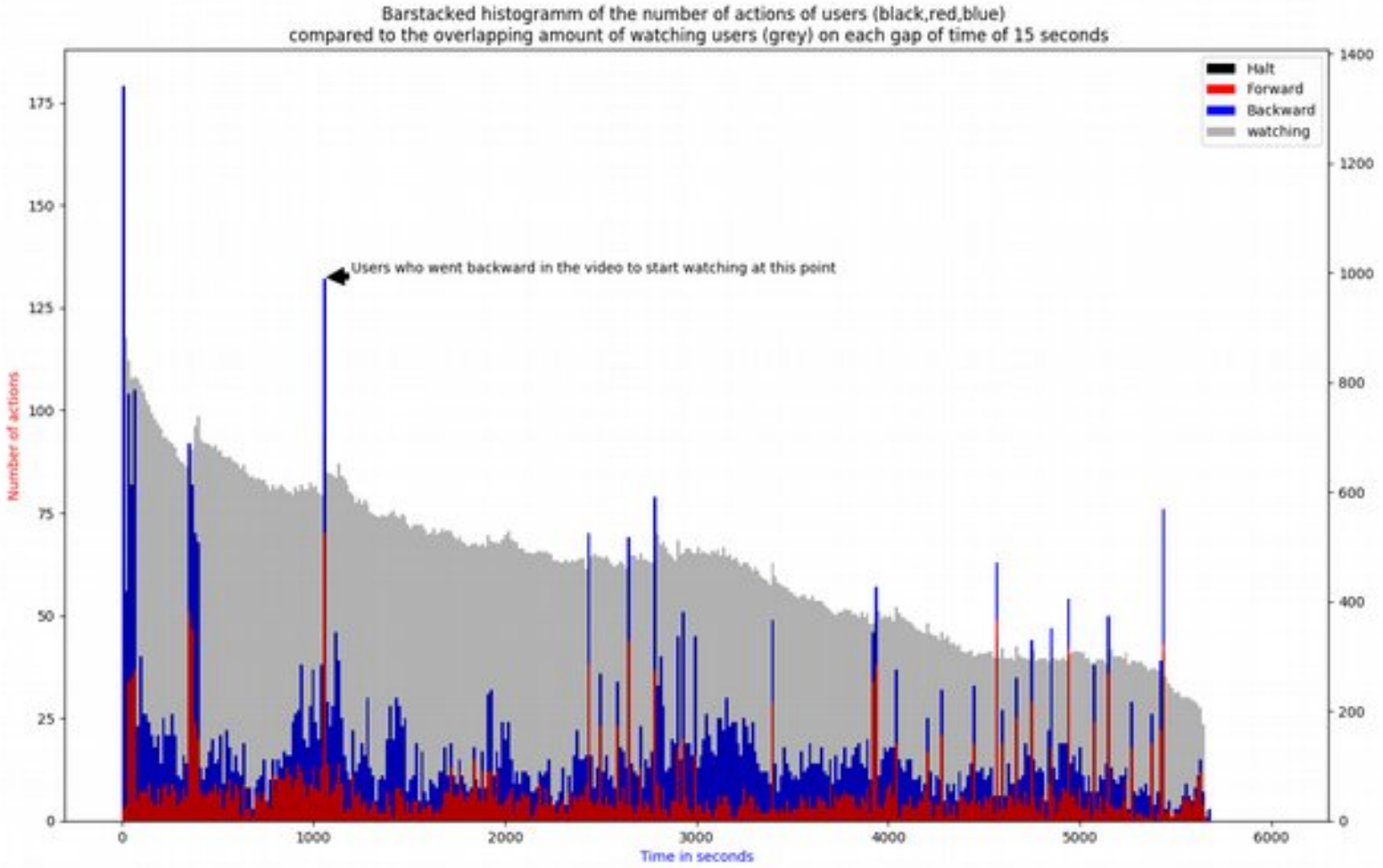
6.4 La connexion temporelle

Travail Marie Humbert—Ropers 2018

Relating video transcriptions and logs

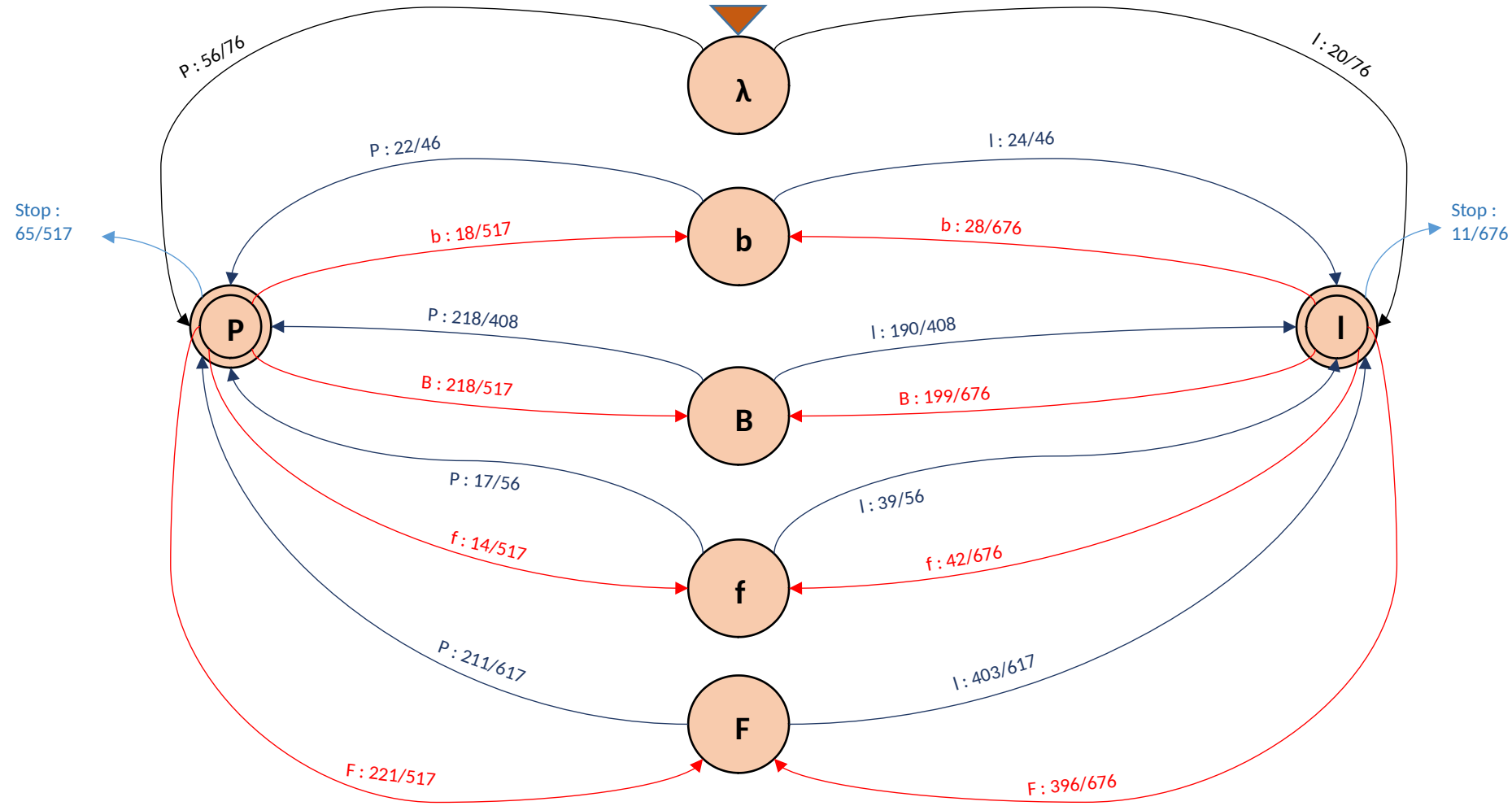
- The key question was: this user stopped after 3mn17s. Do we know why? Can we use the transcriptions to find out?


Break
Forward
Backward
Slides
Number of actions
Time in seconds
0
25
50
75
100
125
150
175
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1000
2000
3000
4000
5000
6000


Barstacked histogramm of the number of actions of users (black,red,blue)
compared to the overlapping amount of watching users (grey) on each gap of time of 15 seconds
Halt
Forward
Backward
watching
Users who went backward in the video to start watching at this point
Number of actions
Time in seconds
0
25
50
75
100
125
150
175
200
400
600
800
1000
1200
1400
1000
2000
3000
4000
5000
6000

Given one video, what are the user actions over that video?

- Alphabet is
  - P
  - B
  - B
  - F
  - F
- K-Testable machines are built


λ
P
I
b
B
f
F
P : 56/76
I : 20/76
P : 22/46
I : 24/46
Stop :
65/517
Stop :
11/676
b : 18/517
b : 28/676
P : 218/408
I : 190/408
B : 218/517
B : 199/676
P : 17/56
I : 39/56
f : 14/517
f : 42/676
P : 211/617
I : 403/617
F : 221/517
F : 396/676