General Presentation and Class Structure

MPRI 2.26.2: Web Data Management

Antoine Amarilli

Friday, December 7th

Web Data Management

- A class about the Web and the data that it contains
- Strong practical aspects but connections to theory
    - e.g., XPath (practice) vs tree automata (theory)
    - e.g., SPARQL (practice) vs regular path queries (theory)

→ A way to see some practice within the confines of MPRI

→ A way to see some exotic theory motivated by practice

Teachers

Antoine Amarilli

Télécom ParisTech

https://a3nm.net/

a3nm@a3nm.net

Pierre Senellart

École normale supérieure

http://pierre.senellart.com/

pierre@senellart.com

Class time and modalities

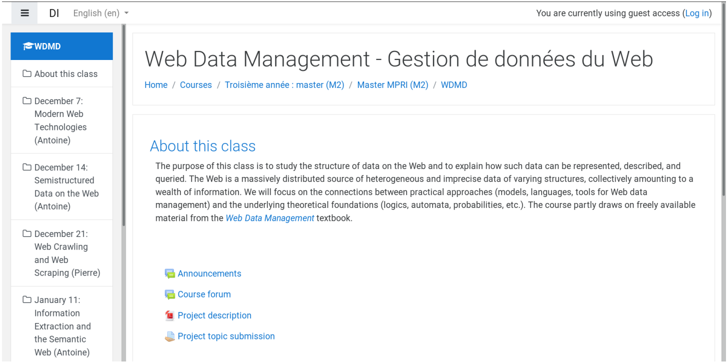
- On Friday afternoon from 16:15 to 19:30 with break(s).
    - Sorry about your weekend plans...

Class time and modalities

- On Friday afternoon from 16:15 to 19:30 with break(s).
    - Sorry about your weekend plans...
- Attendance is not mandatory but we have attendance sheets
    - So we can know whether you sometimes show up in class...

Moodle

`https://moodle.di.ens.fr/course/view.php?id=9` (cf wikimpri)



≡ DI English (en)
You are currently using guest access (Log in)
WDMD
About this class
December 7: Modern Web Technologies (Antoine)
December 14: Semistructured Data on the Web (Antoine)
December 21: Web Crawling and Web Scraping (Pierre)
January 11: Information Extraction and the Semantic Web (Antoine)

Web Data Management - Gestion de données du Web

Home / Courses / Troisième année : master (M2) / Master MPRI (M2) / WDMD

About this class

The purpose of this class is to study the structure of data on the Web and to explain how such data can be represented, described, and queried. The Web is a massively distributed source of heterogeneous and imprecise data of varying structures, collectively amounting to a wealth of information. We will focus on the connections between practical approaches (models, languages, tools for Web data management) and the underlying theoretical foundations (logics, automata, probabilities, etc.). The course partly draws on freely available material from the *Web Data Management* textbook.

- Please **register to the class** on Moodle!
  - **ENS students** can directly use the **SPI CAS** to login
  - **Other students** can create an account
  - **Everyone** can **self-enrol** to the class

What is Moodle good for?

- Finding the class material (slides, etc.)
- Ask questions (better than via email)
- Read questions asked by others
    - You can subscribe to notifications if you wish
- Submit your project (more soon)

Class evaluation

- 50% of the grade will be an exam (on March 1st)
  - → This is required by MPRI rules...

Class evaluation

- 50% of the grade will be an exam (on March 1st)
  - → This is required by MPRI rules...
- 50% of the grade will be a project
  - → Namely...

About that project...

- All details are on Moodle, here are the key points:
- 1 student or 2 students per project
- Free choice of topic related to the Web
- Deadlines and deliverables:

About that project...

- All details are on Moodle, here are the key points:
- 1 student or 2 students per project
- Free choice of topic related to the Web
- Deadlines and deliverables:
  - Dec 21: submit on Moodle the project description and group

About that project...

- All details are on Moodle, here are the key points:
- 1 student or 2 students per project
- Free choice of topic related to the Web
- Deadlines and deliverables:
    - Dec 21: submit on Moodle the project description and group
    - The codebase should be open-source on a public repository
    - There should be a README with minimal documentation

About that project...

- All details are on Moodle, here are the key points:
- 1 student or 2 students per project
- Free choice of topic related to the Web
- Deadlines and deliverables:
  - Dec 21: submit on Moodle the project description and group
  - The codebase should be open-source on a public repository
  - There should be a README with minimal documentation
  - Feb 22: End of project, defense with slides and a demo

About that project...

- All details are on Moodle, here are the key points:
- 1 student or 2 students per project
- Free choice of topic related to the Web
- Deadlines and deliverables:
  - Dec 21: submit on Moodle the project description and group
  - The codebase should be open-source on a public repository
  - There should be a README with minimal documentation
  - Feb 22: End of project, defense with slides and a demo

→ Use the project for...

  - Trying out some original idea
  - Scratching a personal itch
  - Contributing to an existing codebase

About that project...

- All details are **on Moodle**, here are the key points:
- **1 student** or **2 students** per project
- **Free choice of topic** related to the Web
- Deadlines and deliverables:
  - **Dec 21:** submit on Moodle the project description and group
  - The **codebase** should be open-source on a **public repository**
  - There should be a **README** with minimal documentation
  - **Feb 22:** End of project, **defense** with slides and a **demo**

→ Use the project for...

  - Trying out some **original idea**
  - Scratching a **personal itch**
  - Contributing to an **existing codebase**

→ Try to **have fun**! ;-)

Class schedule

- **December 7:** Modern Web Technologies (Antoine)

Class schedule

- December 7: Modern Web Technologies (Antoine)
- December 14: Semistructured Data on the Web (Antoine)

Class schedule

- **December 7:** Modern Web Technologies (Antoine)
- **December 14:** Semistructured Data on the Web (Antoine)
- **December 21:** Web Crawling and Web Scraping (Pierre)

Class schedule

- **December 7:** Modern Web Technologies (Antoine)
- **December 14:** Semistructured Data on the Web (Antoine)
- **December 21:** Web Crawling and Web Scraping (Pierre)
- (Merry Christmas and Happy New Year!)

Class schedule

- **December 7:** Modern Web Technologies (Antoine)
- **December 14:** Semistructured Data on the Web (Antoine)
- **December 21:** Web Crawling and Web Scraping (Pierre)
- (Merry Christmas and Happy New Year!)
- **January 11:** Information Extraction and the Semantic Web (Antoine)

Class schedule

- **December 7:** Modern Web Technologies (Antoine)
- **December 14:** Semistructured Data on the Web (Antoine)
- **December 21:** Web Crawling and Web Scraping (Pierre)
- (Merry Christmas and Happy New Year!)
- **January 11:** Information Extraction and the Semantic Web (Antoine)
- **January 18:** Veracity and Explainability on the Web (Antoine)

Class schedule

- **December 7:** Modern Web Technologies (Antoine)
- **December 14:** Semistructured Data on the Web (Antoine)
- **December 21:** Web Crawling and Web Scraping (Pierre)
- (Merry Christmas and Happy New Year!)
- **January 11:** Information Extraction and the Semantic Web (Antoine)
- **January 18:** Veracity and Explainability on the Web (Antoine)
- **February 1:** Web Information Retrieval (Pierre)

Class schedule

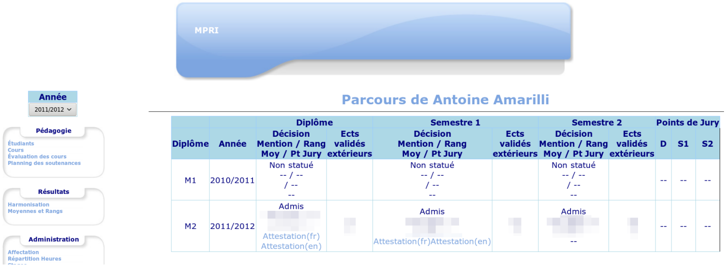
- **December 7:** Modern Web Technologies (Antoine)
- **December 14:** Semistructured Data on the Web (Antoine)
- **December 21:** Web Crawling and Web Scraping (Pierre)
- (Merry Christmas and Happy New Year!)
- **January 11:** Information Extraction and the Semantic Web (Antoine)
- **January 18:** Veracity and Explainability on the Web (Antoine)
- **February 1:** Web Information Retrieval (Pierre)
- **February 8:** Computation and Data Storage at Web Scale (Pierre)

Class schedule

- **December 7:** Modern Web Technologies (Antoine)
- **December 14:** Semistructured Data on the Web (Antoine)
- **December 21:** Web Crawling and Web Scraping (Pierre)
- (Merry Christmas and Happy New Year!)
- **January 11:** Information Extraction and the Semantic Web (Antoine)
- **January 18:** Veracity and Explainability on the Web (Antoine)
- **February 1:** Web Information Retrieval (Pierre)
- **February 8:** Computation and Data Storage at Web Scale (Pierre)
- **February 15:** Web Data Integration, the Deep Web (Pierre)

An MPRI disclaimer

I have been **in your shoes**, not so long ago...



MPRI

Année

2011/2012

Pédagogie

Étudiants
Cours
Évaluation des cours
Planning des soutenances

Résultats

Harmonisation
Moyennes et Rangs

Administration

Affectation
Répartition Heures

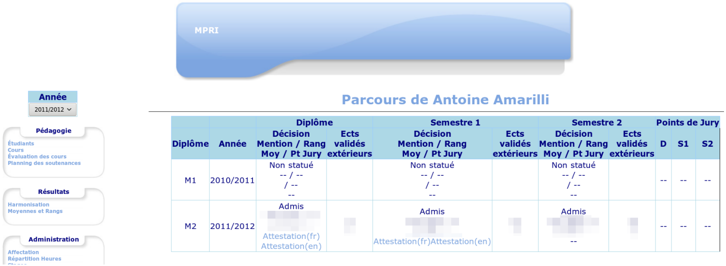
Parcours de Antoine Amarilli

		Diplôme		Semestre 1		Semestre 2		Points de Jury		
Diplôme	Année	Décision Mention / Rang Moy / Pt Jury	Ects validés extérieurs	Décision Mention / Rang Moy / Pt Jury	Ects validés extérieurs	Décision Mention / Rang Moy / Pt Jury	Ects validés extérieurs	D	S1	S2
M1	2010/2011	Non statué -- / -- / -- --		Non statué -- / -- / -- --		Non statué -- / -- / -- --		--	--	--

M2	2011/2012	Admis
		[redacted]
		Attestation(fr)
		Attestation(en)
			[redacted]	Admis
				[redacted]
				Attestation(fr)Attestation(en)
					[redacted]	Admis
						[redacted]
						--
							[redacted]	--	--	--

An MPRI disclaimer

I have been **in your shoes**, not so long ago…



MPRI

Année

2011/2012

Pédagogie

Étudiants
Cours
Évaluation des cours
Planning des soutenances

Résultats

Harmonisation
Moyennes et Rangs

Administration

Affectation
Répartition Heures

Parcours de Antoine Amarilli

		Diplôme		Semestre 1		Semestre 2		Points de Jury		
Diplôme	Année	Décision Mention / Rang Moy / Pt Jury	Ects validés extérieurs	Décision Mention / Rang Moy / Pt Jury	Ects validés extérieurs	Décision Mention / Rang Moy / Pt Jury	Ects validés extérieurs	D	S1	S2
M1	2010/2011	Non statué -- / -- / -- --		Non statué -- / -- / -- --		Non statué -- / -- / -- --		--	--	--

M2	2011/2012	Admis
		[redacted]
		Attestation(fr)
		Attestation(en)
			[redacted]	Admis
				[redacted]
				Attestation(fr)Attestation(en)
					[redacted]	Admis
						[redacted]
						--
							[redacted]	--	--	--

What I remember from these days is not great…

- **Ultra-specialized** classes
- **No effort** to teach prerequisites
- Only relevant to people who want to **specialize** in the field
- Only **theory** and no **practice**

An MPRI confession

Now I'm a teacher and understand why teachers teach like that:

- They enjoy research more than teaching
- They are promoted based on research not teaching
- They are here to find PhD students to do more research
- They are short on time so…
    - Making complicated things understandable takes time
    - It's far easier to recycle existing slides about stuff you know!

Will this class be any different?

- Less incomprehensible theory and more shallow practice
- The project can be fun (hopefully)
- The class material/structure is probably not perfect, sorry...
- OK I won't try to hide it: we are hiring!

Will this class be any different?

- Less incomprehensible theory and more shallow practice
- The project can be fun (hopefully)
- The class material/structure is probably not perfect, sorry...
- OK I won't try to hide it: we are hiring!

Will this class be any different?

- Less incomprehensible theory and more shallow practice
- The project can be fun (hopefully)
- The class material/structure is probably not perfect, sorry...
- OK I won't try to hide it: we are hiring!

- ... but most of this class isn't so related to what I do, so...