

# General Presentation and Class Structure

MPRI 2.26.2: Web Data Management

---

Antoine Amarilli



# 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 Paris



[a3nm.net](http://a3nm.net)

[a3nm@a3nm.net](mailto:a3nm@a3nm.net)

Liat Peterfreund

CNRS, Paris-Est University



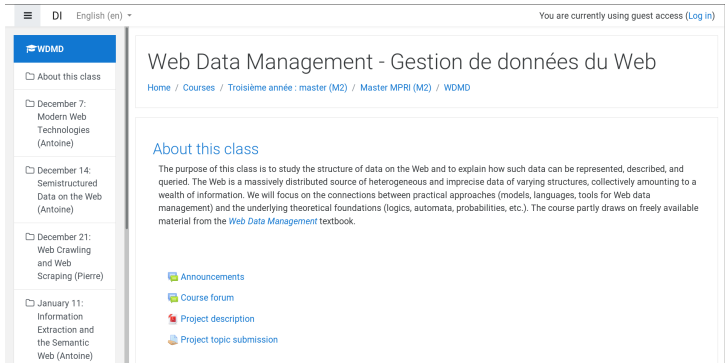
[sites.google.com/view/liatpeterfreund](https://sites.google.com/view/liatpeterfreund)

[liatpf.cs@gmail.com](mailto:liatpf.cs@gmail.com)

# Class time and modalities

- On **Monday afternoon** from **16:15** to **19:15**
- Attendance is **not mandatory** but we have **attendance sheets**
  - So we can know whether you sometimes show up in class...

<https://moodle.r2.enst.fr/moodle/enrol/index.php?id=142>



The screenshot shows a Moodle course page. At the top, there is a navigation bar with 'DI English (en)' and 'You are currently using guest access (Log in)'. The course title is 'Web Data Management - Gestion de données du Web'. Below the title is a breadcrumb trail: 'Home / Courses / Troisième année : master (M2) / Master MPRI (M2) / WDMO'. The main content area is titled 'About this class' and contains a paragraph of text: '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.' Below the text are four links: 'Announcements', 'Course forum', 'Project description', and 'Project topic submission'. On the left side, there is a sidebar with a list of course activities: '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)', and 'January 11: Information Extraction and the Semantic Web (Antoine)'.

- Linked from the **MPRI Wiki**
- Polytechnique and Paris-Saclay students can log in directly, otherwise I will create an account for you
- Once you have an account, you can **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 6th)
  - This is required by MPRI rules...

# Class evaluation

- **50%** of the grade will be an **exam** (on March 6th)
  - 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 14:** 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 14:** 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 14:** 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
  - **February 27:** 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 14:** 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
    - **February 27:** 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 14:** 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
    - **February 27:** 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 dates

- December **5** and **12** (Antoine)
- Christmas break: No class from Dec 19 to Jan 9
- January **9, 16, 23, 30** (Liat)
- February **6** and **13** (Antoine)
- Project defenses: **February 27**
- Exam: **March 6**

# Class topics

- Modern Web technologies (today)
- Semistructured data
- Web crawling and scraping
- Web information retrieval (search engines)
- Computation and data storage at web scale (Hadoop, Spark...)
- NoSQL
- Information extraction
- Semantic Web
- Uncertainty in databases



# An MPRI disclaimer

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

MPRI

## Parcours de Antoine Amarilli

**Année**  
2011/2012 ▾

**Pédagogie**  
Étudiants  
Cours  
Évaluation des cours  
Planing des soutenance

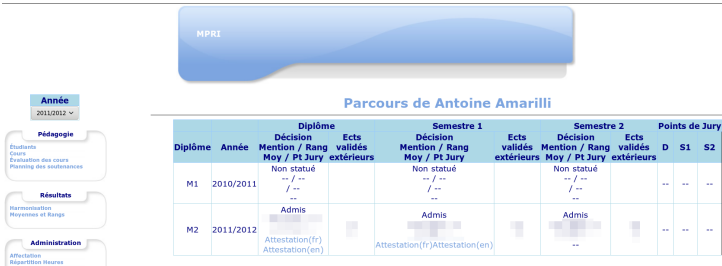
**Résultats**  
Harmonisation  
Moyennes et Rangs

**Administration**  
Affectation  
Répartition Heures

Diplôme	Année	Diplôme		Semestre 1		Semestre 2		Points de Jury		
		Décision Mention / Rang	Ects validés Moy / Pt Jury extérieurs	Décision Mention / Rang	Ects validés extérieurs	Décision Mention / Rang	Ects validés extérieurs	D	S1	S2
M1	2010/2011	Non statué -- / -- / -- --		Non statué -- / -- / -- --		Non statué -- / -- / -- --		--	--	--
M2	2011/2012	Admis		Admis		Admis		--	--	--
		Attestation(fr) Attestation(en)		Attestation(fr)Attestation(en)		--				

# An MPRI disclaimer

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



The screenshot shows the MPRI website interface. At the top, there is a blue header with the text "MPRI". Below it, the title "Parcours de Antoine Amarilli" is displayed. On the left side, there are three main navigation categories: "Année" (Year) with a dropdown menu showing "2011/2012", "Pédagogie" (Pedagogy) with sub-items "Étudiants", "Cours", "Évaluation des cours", and "Passing des soutenance"; "Résultats" (Results) with sub-items "Harmonisation", "Moyennes et Rangs"; and "Administration" (Administration) with sub-items "Affectation" and "Répartition Heures". The main content area is a table titled "Parcours de Antoine Amarilli" showing academic performance for two semesters.

Diplôme	Année	Diplôme		Semestre 1		Semestre 2		Points de Jury		
		Décision Mention / Rang	Ects validés Moy / Pt Jury extérieurs	Décision Mention / Rang	Ects validés extérieurs	Décision Mention / Rang	Ects validés Moy / Pt Jury extérieurs	D	S1	S2
M1	2010/2011	Non statué -- / -- / -- --		Non statué -- / -- / -- --		Non statué -- / -- / -- --		--	--	--
M2	2011/2012	Admis		Admis		Admis		--	--	--
		Attestation(fr) Attestation(en)		Attestation(fr)Attestation(en)		--				

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:

## 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!**

