Funded by
the European Union

ITHACA

Artificial Intelligence to enhance civic participation

TrustworthyAI by design: Insights from the AI, Big Data and Democracy Taskforce

1/10/2024

ITHACA

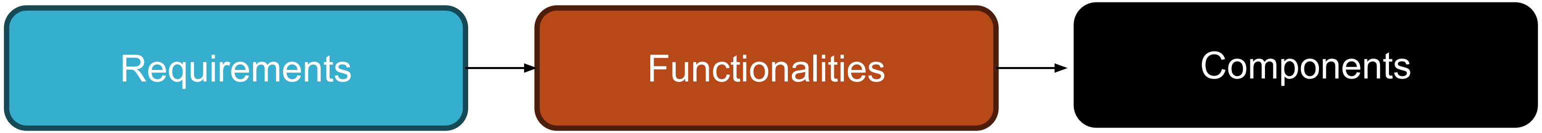
01.

The ITHACA platform

The ITHACA platform Overview

Description: The aim is to develop a trustworthy AI platform to facilitate civic participation in the democratic processes of the project.
Key factors contributing to concerns about trustworthiness in AI per each perspective:

- **Technical:** accurate, robust and provide explainability
- **User:** Prioritise usability, safety and privacy
- **Society:** Legal compliance, uphold ethical standards, ensure fairness, accountability and be environmental conscious



Requirements → Functionalities → Components

The ITHACA platform Overview

STEP 1 – Identify key Technical Requirements based on the project's Use Cases (Technical perspective) and User Requirements (User and Societal perspective)

Requirements extracted through questionnaires and workshops including all relevant stakeholders:

- **Users**
- **Technical experts**
- **Data analysts**
- **Ethical and Legal advocates**

Use Cases defined:

Actors involved	✓
Scenarios	✓
Pre-conditions	✓
Post-conditions	✓
Quality Requirements	✓
Success Criteria	✓

The ITHACA platform Overview

STEP 2 – Identification of the **Functional/ Technical Requirements (65 total)**

Technical Requirements	Identified Functionalities
Requirement 1: Content management	19
Requirement 2: Content viewing and integration of input from other sources	13
Requirement 3: Guidance/ Navigation bot	8
Requirement 4: Accessibility	4
Requirement 5: Content summarization & simplification	4
Requirement 6: Login mechanism	4
Requirement 7: User’s personal page	4
Requirement 8: HMI interface (front-end)	4
Requirement 9: Security	5

The ITHACA platform Overview

STEP 2 – Identification of the **Functional/ Technical Requirements**

Requirement 1: Content management

- Functionality 1: Forum-based functionality
- Functionality 2: Interoperability with Guidance/ navigation bot
- Functionality 3: Quick evaluation of posts
- Functionality 4: Quick respond options
- Functionality 5: App activators
- Functionality 6: Interoperability with other built-in or external apps
- Functionality 7: Moderation of content
- Functionality 8: Content readable by TTS apps
- Functionality 9: Microphone option to input field
- Functionality 10: Basic info incorporated into posts
- Functionality 10: Hidden post information
- Functionality 11: Pre-defined forms and cells
- Functionality 12: Similarity analysis check
- Functionality 13: Posts to include links and embedded formats
- Functionality 14: Live streaming video
- Functionality 15: Customisation of voting posts
- Functionality 16: Providing content feedback
- Functionality 17: Providing platform feedback
- Functionality 18: Conformity check
- Functionality 19: Security applications related to the content

The ITHACA platform Overview

Requirement 2: Content viewing and integration of input from other sources

- Functionality 20: Filtering and ranking options
- Functionality 21: Content presentation to include external sources
- Functionality 20: Newsfeed section
- Functionality 21: Store filtering options
- Functionality 23: Interoperability of presentation/ modification options with guidance/ navigation bot
- Functionality 24: Accessibility modifications
- Functionality 25: Interoperability with summarisation function
- Functionality 26: Provide direct link options to the presented content
- Functionality 27: Additional post info to be included on a side window or within post
- Functionality 28: Live streams presentation
- Functionality 29: TTS function for oral presentation
- Functionality 30: Format of the presented results to include alternative text/ content based on summarisation/ simplification app results
- Functionality 31: Include charts for the presentation of voting/ polls
- Functionality 32: FAQ section

The ITHACA platform Overview

Requirement 3: Guidance/ Navigation (bot)

- Functionality 33: Greeting itself and the platform’s.
- Functionality 34: Tutorial feature
- Functionality 35: Voice commands
- Functionality 36: Enabling platform commands/ actions/ features
- Functionality 37: Keep stored key contact persons information and link them to the users
- Functionality 38: Provide feedback about responses
- Functionality 39: Detection of special needs
- Functionality 40: Detection of external TTS app

Requirement 4: Accessibility

- Functionality 41: Accessibility modifications on the presented content
- Functionality 42: Make use of STT technology
- Functionality 43: Make use of TTS technology
- Functionality 44: Keyboard

Requirement 5: Content summarisation & simplification

- Functionality 45: Summarisation and simplification of text
- Functionality 46: Summarisation and simplification feedback
- Functionality 47: Argument visualisation
- Functionality 48: Translate content

The ITHACA platform Overview

Requirement 6: Login mechanism

- Functionality 49: Creation of personal account
- Functionality 50: Store user’s personal information
- Functionality 51: Store user’s personal preferences
- Functionality 52: Keeping the user logged in

Requirement 7: Personal page

- Functionality 53: Integrated mailing function
- Functionality 54: User performance
- Functionality 55: Modifications and pre-defined preferences
- Functionality 56: Disclaimer about data stored

Requirement 8: HMI interface (Front End)

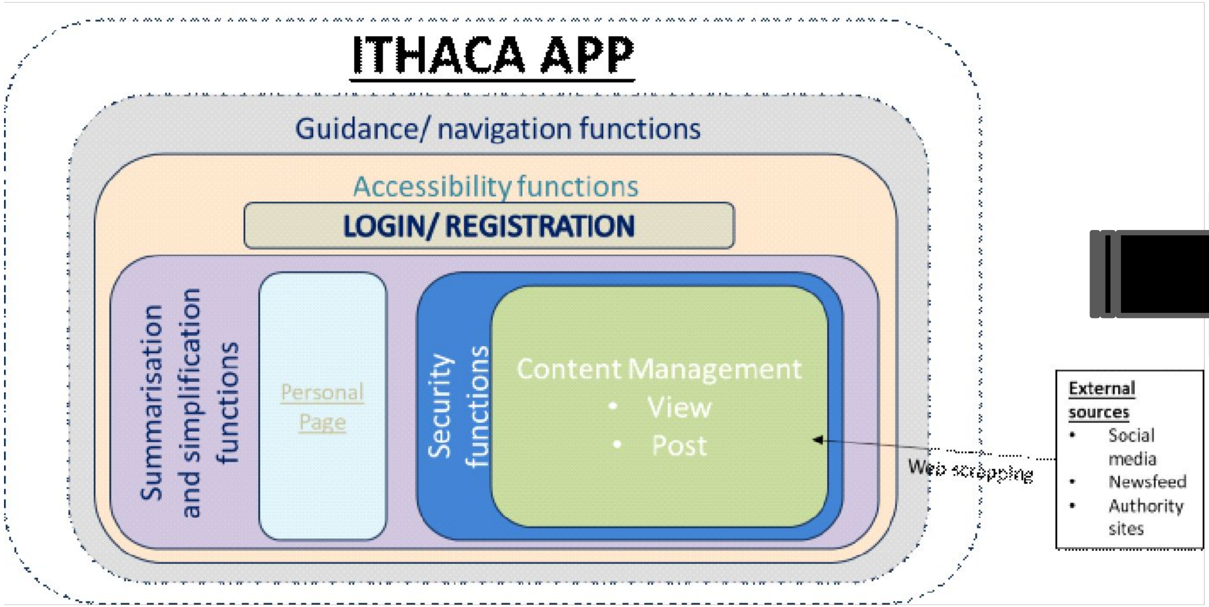
- Functionality 57: Resemblance to other well-known and popular platforms
- Functionality 58: Icons to initiate actions and apps
- Functionality 59: Explainability of actions
- Functionality 60: Interpretation of the platform through voice description

Requirement 9: Security

- Functionality 60: Protection of user’s personal data
- Functionality 62: Security of stored data
- Functionality 63: Inappropriate content
- Functionality 64: External attacks
- Functionality 65: Blurring of inappropriate content

The ITHACA platform Overview

STEP 3 – Identification of the **Functional/ Technical Components**



Architecture Design & Components

02.

The Architecture

The ITHACA platform Architecture

Description:

We proposed a Hybrid Architectures (Microservices combined with Event-Driven Messaging):

•**Modularity**: Hybrid architectures leverage modular design principles, allowing components to be loosely coupled and independently deployable.

•**Scalability**: The combination of Microservices and Event-Driven Messaging enables scalable solutions that can adapt to varying workloads.

Integration of Event-Driven and Microservices Concepts:

•**Event-Driven Communication**: communication through events, allowing microservices to react to changes in the system in real-time.

•**Microservices Independence**: Each microservice retains its autonomy, while event-driven communication facilitates coordination and collaboration.

Operating System Layer

Linux CentOS: The foundation of the application server, providing a stable and secure environment fo

Hosting application services.

Filesystem and Datastore Layer

- **MongoDB**: NoSQL database used for storing application data, offering high performance and scalability.
- Manages storage for application data, configurations, logs, and other filesystem operations.

Infrastructure and orchestration layer

- **Docker Containers**: Services are containerized to ensure consistency across different

environments and ease of deployment.

Middleware Layer

- **Kafka Message Hub**: Acts as a message broker, facilitating communication between

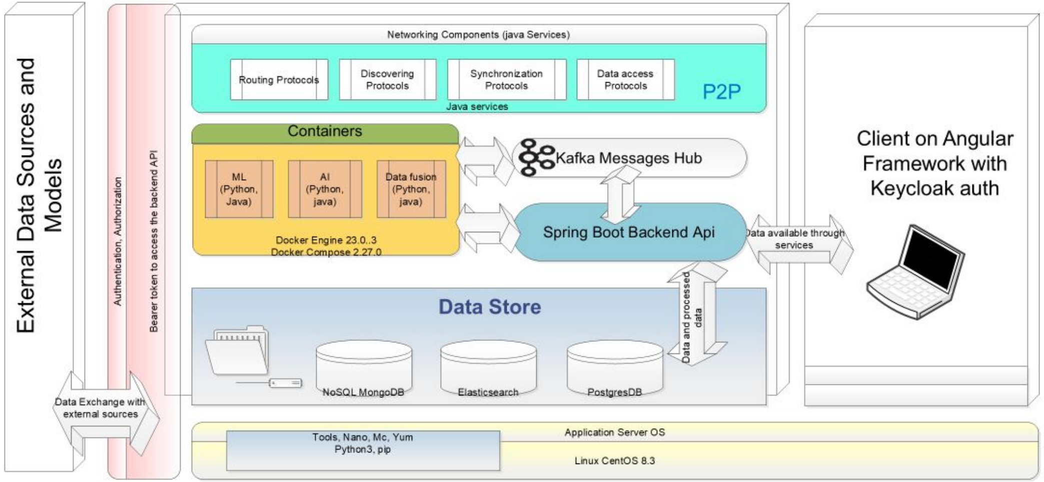
different services and ensuring real-time data streaming and processing.

Application Layer

- **Spring Boot Backend API**: Manages business logic, interacts with the MongoDB datastore, and c
for message processing.

Client Layer

- **Angular Framework**: Frontend application that interacts with the backend via APIs.
- **Keycloak Authentication**: Manages user authentication and authorization, ensuring secure access to the application.



03.

AI Tools for trustworthiness

Trustworthy AI in ITHACA (1/2)

In the context of WP5, the ITHACA partners developed three supportive tools that are applied alongside AI-based civic participation systems to evaluate the latter in terms of Fairness, Security and Privacy:

Tools:

- AI Fairness Tool:
  - Purpose: Conformity with the fairness principle, to promote equality, inclusivity and oppose discrimination as well as render the evaluated AI system more trustworthy and unbiased.
  - Functionality: The tool employs objective fairness metrics (e.g. treatment equality, disparate impact, between group generalized entropy etc.), to evaluate whether an AI toxic censoring model associates people from vulnerable groups with offensive language with a higher probability, thus, being biased and unfair.

- PPML – Privacy Preserving Machine Learning :
  - Purpose: Application of data concealment techniques to safeguard user privacy.
  - Functionality: The tool utilizes a Differential Privacy method to perturb sensitive user data by adding random noise, so that it would not be possible for a potentially malicious actor to infer any personal, potentially identifying information in case of a potential attack via specifically engineered inputs to the model.

Trustworthy AI in ITHACA (2/2)

In the context of WP5, the ITHACA partners developed three supportive tools that are applied alongside AI-based civic participation systems to evaluate the latter in terms of Fairness, Security and Privacy:

Tools:

- AI Cybersecurity :
  - Purpose: detection of possible security breaches and threats within AI models
  - Functionality: The open-source tool named ModelScan is proposed as the AI Cybersecurity tool. This tool scans for malicious code that may impose security vulnerabilities (e.g. complete control from malicious actors) in AI systems.

- Visual Component :
  - Purpose: The Visual Component is intended for a responsible party (i.e. human moderator or platform maintainer/engineer) to have control over big data and events arising from the function of the above evaluation tools.
  - Functionality:
    - Highlighting words (explainability mechanism) heavily associated with toxic posts by the AI Fairness tool with a color from a palette from green to red indicating their level of toxicity
    - Visualizing metrics such as Attacker Advantage to evaluate the level of privacy that the PPML tool offers, while the authors of posts depicted in the visual component remain anonymous to maintain impartiality, fairness, and privacy.

Thank you!

Aristotelis Spiliotis: aspiliotis@certh.gr
Adrian Dragota: Adrian.Dragota@simavi.ro
Iliana Loi: loi@ceid.upatras.gr

ITHACA
AI To Enhance Civic Participation