



# Accessible Extended Reality and the Immersive Web

**Open Standards and Open Source**

Dave Raggett W3C/ERCIM

[dsr@w3.org](mailto:dsr@w3.org)

June 2024



# Many Applications

*virtual and augmented reality*



## **Shopping, Entertainment, Education, Industry, Online meetings and Desktop replacement:**

- Smart phones and tablets can be used to show how your new kitchen, or new furniture and decor for your living room would look as an augmented reality experience
- See how you look in a smart mirror (e.g. your phone) when wearing some new clothes, new shoes, different hair styles, jewellery and glasses, before purchasing them online
- Play VR and AR games for mutual enjoyment with your friends
- For industry, smart glasses for machine maintenance, smart warehouses, embracing the IoT
- Online meetings that project you and your companions into a shared immersive VR or AR environment
- VR and AR will function as a computer desktop replacement, offering unlimited virtual screens via headsets or smart glasses



# Digital Twins

*virtual and augmented reality for the IoT*



- Digital twins are virtual embodiments of physical systems, processes and people, e.g. sensors and actuators in the real world
- Digital twins are part of an extended reality that allow you to interact with things whether they are real or imagined
- Digital twins are associated with digital footprints that are visible in extended reality, e.g. using smart glasses\*



- Semantic descriptions, e.g. a thermostat for a particular room using Celsius for units
- Metadata such as product name, serial number, installation date, owner, test certificates, ...
- Interaction affordances, e.g. as virtual objects with properties, actions and events
- Digital twin avatars for repair instructions

\* Using smart glasses to view and interact with information about machines and processes



# The Immersive Web



## **Augmented Reality (AR)**

Interactive experience combining real world and computer-generated content, e.g. visual, auditory, haptic

## **Virtual Reality (VR)**

Interactive experience in a computer generated environment

## **Extended Reality (XR)**

A term embracing both augmented and virtual reality

## **Metaverse**

Virtual worlds (or universes) in which users are represented by digital avatars and interact according to the rules of that world, which may depart significantly from the real world

## **Omniverse**

The superset of all universes, both real and imagined

## **Immersive Web**

Open Web standards for extended reality experiences embracing the omniverse



# Accessible Extended Reality

*enabling everyone to use applications*



- Based upon open standards and community driven open-source implementations
- Scalable across different devices
- Including head mounted displays smart glasses\* and regular 4K UHD screens
- Client and server-side rendering
- Peer to peer networking for low latency interaction
- Privacy preserving accessibility based upon high-level models
- Taxonomy of resources and behaviours
- Scripting – enabling adaptation to device and accessibility requirements
- Building on top of established standards for protocols, formats and APIs, including WebXR

\* Metalenses will enable comfortable lightweight smart glasses



# Both client and server based rendering



## Client-side Rendering

- Pre-loading libraries of resources for better performance
- Substitution or adaptation of similar resources as needed
- Local processing on webcam, key strokes, mouse input etc.
- Use of GPU for rendering
- WebRTC for peer to peer

## Server-side Rendering

- Faster as no need to download large resources
- Server streams video and audio to browser
- Browser streams interaction to server having done local processing on webcam, key strokes, mouse input etc.



# Taxonomies of Resources



- Knowledge-based approach
  - Knowledge graphs + links
- Models of resources and associated behaviours
- Multiple levels of abstraction
  - High level, e.g. building, animal, ...
  - Medium level with exposed API
  - Low level with 3D data, textures
- Links to lower level resources, e.g. 3D models, texture tiles, audio/video files, scripts
- Resources have URNs so they can be found in named libraries
  - Improved caching performance via resource sharing
  - Similarity metrics for use in resource substitution and potential adaptation
- Accessibility based upon multiple levels of abstraction
- Intent based interaction rather than relying on low-level UI events
- Reduce need to share user preferences with applications

*Intent-based interaction decouples applications from details of user input, enabling users to select the best means for their own needs*



# Simple Descriptions



- To ease authoring effort, the taxonomy supports lightweight descriptions
- Using linked scripts, these are expanded into detailed models
- An example is a building with
  - Floor plan, doors, stairs, windows
  - Properties as annotations, e.g. wallpaper, picture frames, ...
- Fractal-like models generated from simple descriptions
- Using pseudo-random number generators with given seeds
- Generated identifiers, e.g. for rooms, can be used to retrieve properties as annotations
  - What's on the table, walls, bookshelves, etc.

Further out, use of generative AI with simple text prompts for artistically themed environments



# Behaviours



*intent*: an aim, purpose, goal or objective

- Intents include picking something up, walking to a new location, opening a door, etc.
- The intent's behaviour may be implicit, e.g. to go upstairs, you have to walk up the stairs
- This allows intents to be treated as goals leaving their realisation to scripts to resolve in a transparent and expected way
- Intents may be implicit given the context and current goals
- Named behaviours as part of taxonomies
  - Stationary, walking, trotting, galloping (for horses)
- APIs for invoking behaviours
  - Described in the taxonomy akin to actions in web of things
- High and low level behaviours
  - Named intents map to high level behaviours, which in turn map to ...
  - Low level behaviours, e.g. skeletal movements and skin models

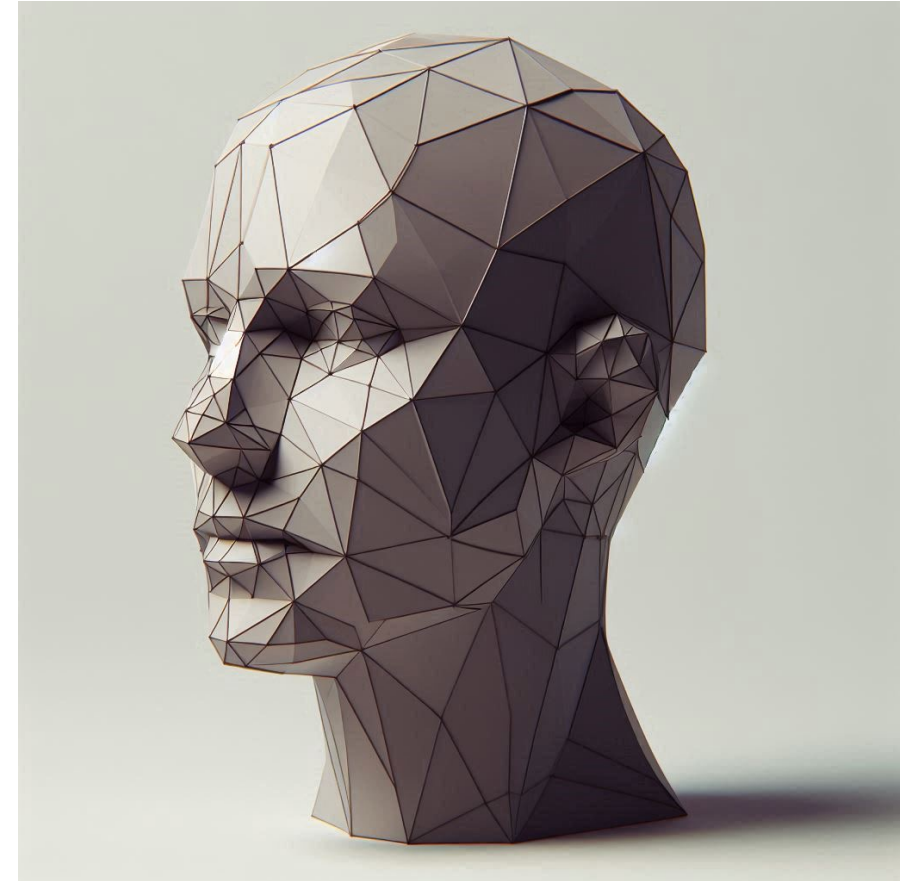


# Animating Avatars



- An avatar is how you are shown in an extended reality context
- It could look like you, or it could be something very different
  - It's your personal preference
- But it should speak like you, and if possible, show your facial expressions as you speak
- Your webcam can be used in real-time to fit a simple wireframe model to your face along with detecting your expressions and gaze\*
- Wireframe is textured with warped video tiles or cached library of expressions
- Other parts synthesised as needed, e.g. back of head
  - Customisable avatars composed from catalogue of models, e.g. hair, head, face, body, clothes, shoes, accessories
- Behavioural models for natural movements

\* Using a parameterised wireframe model that is adapted to each video frame using either CNNs or variations on the Viola-Jones algorithm for face detection. Note the potential for exploiting the WebNN API for hardware acceleration.





# Networking and Discovery



- Efficient communication between agents that are locally visible to each other\*
  - For example in the same VR room
- Peer to peer via WebRTC
  - Streaming media
  - Message exchange
- Reliant on server for setup
- Scaling challenges as agents appear and disappear
- Entering a room or space enables others to discover you
- Privacy preferences covering how much personal information is disclosed
- Context descriptions are given at multiple levels of abstraction
  - e.g. list of people in the room
- Searchable using taxonomic info
  - e.g. people, tables and chairs

\* For scalability, fog can make distant objects invisible for open scenes

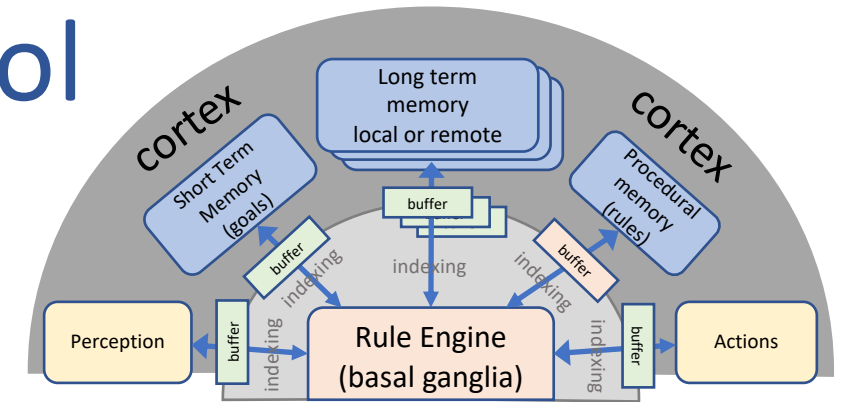


# Cognitive Low-Code Control

- Providing application developers with the option to use a low-code environment for specifying real-time behaviour
- Event-driven concurrent threads of behaviour using APIs exposed by resources as described in taxonomies
- Easy to learn, convenient syntax for chunks\* and condition-action rules
  - W3C Cognitive AI CG's [Chunks & Rules](#) specification
  - Higher level and easier to use than RDF
- Mature open-source JavaScript library
- Extension to distributed agents, e.g. swarms via asynchronous message exchange

\* Chunks are sets of properties, i.e. name/value pairs, corresponding to a set of RDF triples with same subject

## Cognition – Sequential Rule Engine



- The cortex holds a set of modules, each of which is associated with a buffer that holds a single chunk
  - Predefined operations on buffers in analogy with REST
- Inspired by John Anderson's ACT-R
  - Mimics characteristics of human cognition and memory, including spreading activation and the forgetting curve
  - Rule conditions and actions specify which cognitive module buffer they apply to
  - Variables are scoped to the rule they appear in
  - Actions either directly update the buffer or invoke operations on the buffer's cortical module, which asynchronously updates the buffer
  - Extensible suite of cortical operations
  - Perception builds live models of the environment including events that trigger corresponding behaviours
  - Actions expressed as intents to be realised as appropriate
  - Reasoning decoupled from real-time control over external actions, e.g. a [robot arm](#)



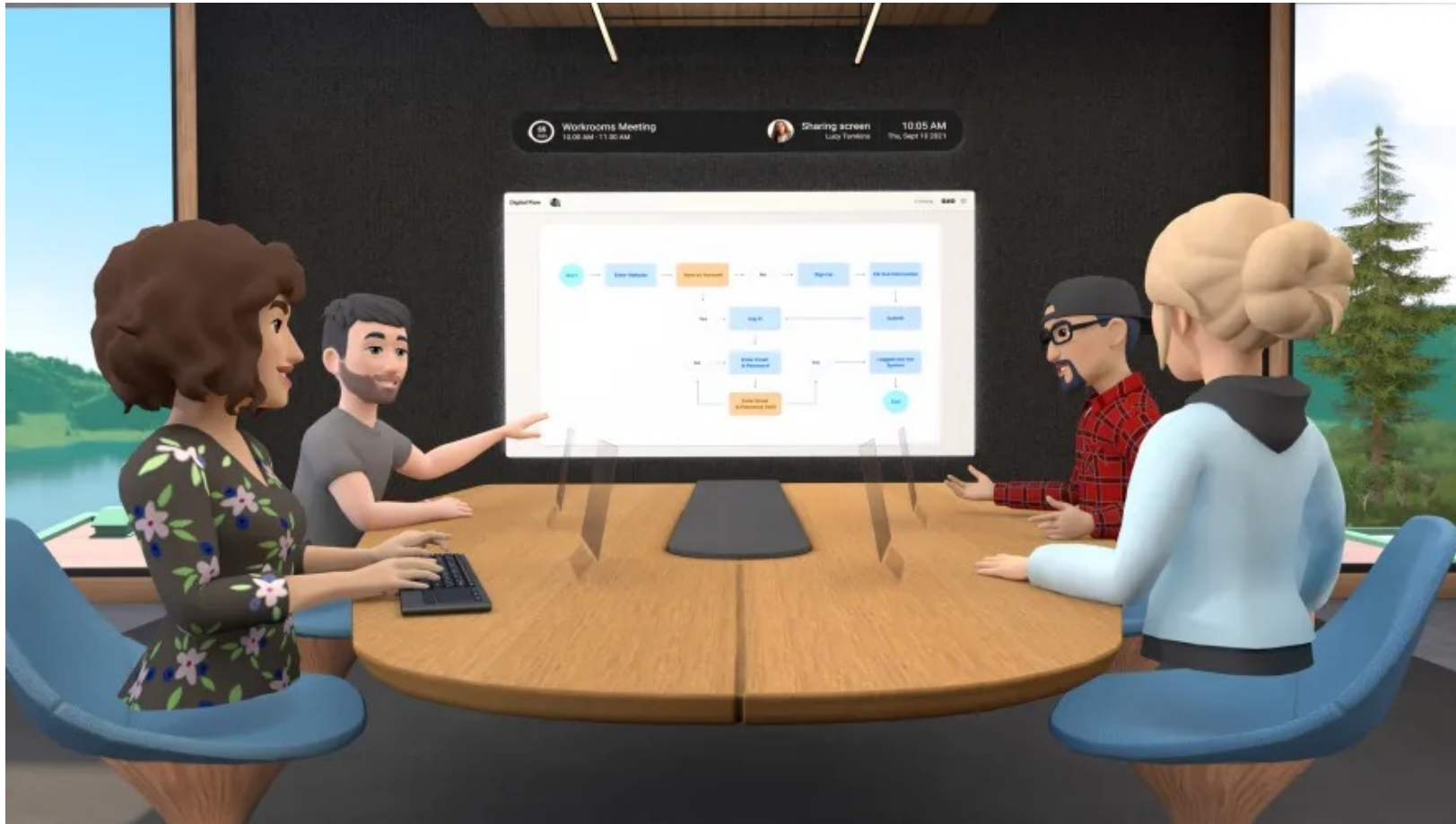
# Microsoft's metaverse



- [Microsoft Mesh](#) offers immersive meetings as an extension to Microsoft Teams
- You can design your avatars including flesh tones and clothes
- The avatars only exist from the waist up!
- You can build digital twins e.g. a hospital ship or office campus
- Proprietary solution



# Meta's metaverse



- Immersive [meeting rooms](#) similar to Microsoft Mesh
- Design your avatars from suite of options
- The avatars only exist from the waist up!
- Spatial audio
- Shared whiteboard
- Hand gestures for control
- Proprietary solution



# Are Metaverses Ready For Everyday Use?



- Higher quality rendering is needed to reduce meeting fatigue
- Current avatars are not intended to be life-like copies of real people
  - Do they adequately mimic your gaze and facial expression?
- Are current computers fast enough to capture and mimic detailed facial expressions in real-time?
  - Real-time automatic subtitles for people with hearing impairments
  - Speech synthesis for people with speech impairments
  - Speech/text understanding for real-time gesture generation
- Full body avatars rather than waist up
- This requires full body modelling with natural movements, e.g. walking, sitting, turning, pointing and other gestures ...
  - Including head and hand gestures
  - Natural movements involves training generative models on real data for people's movements
  - Movement generated on executing intents
- Need for open standards rather than proprietary solutions
- Using web browser rather than proprietary apps



# Developmental Challenges



- Lightweight graph format for taxonomies and applications
  - Chunks would be good choice and can layer on top of RDF
  - Embedded links to external resources
- Developing real-time software for detecting gaze, facial expressions and fitting and texturing a simple wireframe model\*
- Demonstrating feasibility and scalability using web browsers
  - WebGPU, WebNN, WebXR, WebRTC, WASM ...
- Evangelising the vision to build a vibrant community of users, developers and websites
- Selecting a few use cases to show case the potential
- Integration with generative AI
- Enabling a viable open ecosystem with sound business models and strong competition
- Advertising or paywalls or free: how to achieve a balance?

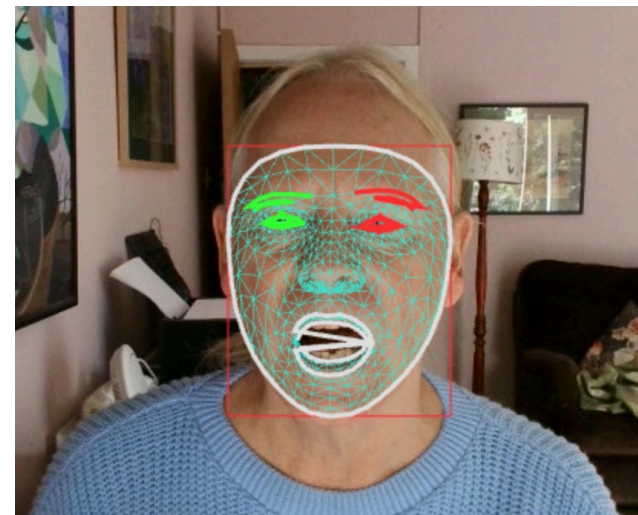
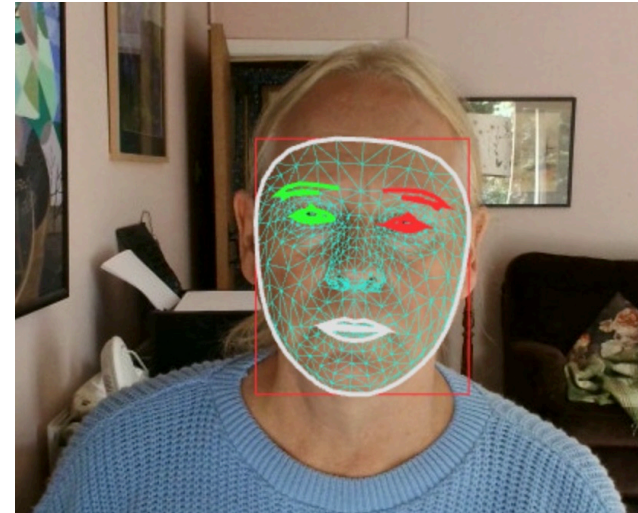
\* Volunteers sought for help with working on open source solution for web browsers



# Facial Mesh Fitting



- [TensorFlow.js](#) includes support for recognising facial landmarks and a [web-based demo](#) that renders this over webcam video at > 60 FPS
- It works well with or without glasses, but there is a little jitter between frames
- It does a good job at fitting an open mouth and smiling, but fails to track my eyebrows when I raise them
- Doesn't work when you bend your head down a lot, presumably as this wasn't in the training data
- The main client-side script is 98 thousand lines!





# Potential Improvements

*Can we do any better?*



- A simpler wireframe mesh will reduce the computational cost, leaving more time for other work and offering better scaling, i.e. more people in the scene
- Use video to extract texture tiles for high quality rendering to compensate for simpler mesh
- Eyebrow movements handled via textures rather than mesh changes
- Your hair and clothes are matched to latent models to generate lifelike avatars
- This further requires pose detection, using modelling for hidden parts, e.g. your legs when sitting behind a table
- The laptop webcam can only see your face and shoulders, and another camera would be needed to observe hand gestures if needed for control
- The body pose is automatically generated when you execute an intent to move, e.g. to another location
- Of course in real-life you remain seated in front of your laptop or holding your smart phone



# Research Ideas



- Use an existing facial landmark model to bootstrap a much simpler one
  - Directly fit 3D model parameters\*
- Generate the training data by adapting an existing dataset
  - Plus photogrammetry on images captured from browser's webcam
- Exploit WebNN and WebGPU for hardware acceleration
- Clean, lean codebase rather than huge bloated libraries
- New work on clothing and pose models for webcam video
- Web-based data capture from volunteers
  - Privacy friendly federated machine learning in the web browser
- Similar effort needed for modelling natural body movements
  - Video captured by smart phones
  - Video from movies, tv dramas?
- Work on taxonomies and models for AR/VR environments
  - Including generative scripts

\* Simplified universal parameterised 3D mesh model that fits all people



# 3D Face Models



- Need 3D models for rendering avatars from different view points, e.g. those of other people in a meeting room
- One approach would be to create a web application to create 3D face models using photogrammetry on webcam images captured by the browser
- Ask volunteers to turn their head to left, to centre, to right, and tilted upwards and downwards
- Potential for privacy-friendly federated learning avoiding uploading images
- Stage 1: train system to compute 3D mesh models from set of images with different poses
- Stage 2: fit each mesh with universal parameterised 3D model
- Stage 3: generate dataset of images + 3D model parameters
- Stage 4: train CNN to predict model parameters from single images
- Stage 5 test and tune application to fit 3D model and extract texture tiles from video frames



# Next Steps



- Recruit partners interested in launching a proposed **W3C Accessible Extended Reality Omniverse (AERO) CG\*** and active participation in associated open source projects, see:
  - [W3C Community Group Process](#)
- Coordination with [W3C Immersive Web CG](#)
  - Focussed on WebXR API, see [explainer](#)
  - Not working on VR/AR browsers and building “The Metaverse”
- Coordination with [XR Access](#)
  - See [Accessible Development for XR](#)
- Coordination with [WAI Accessible Platform Architectures \(APA\)](#)
  - See WG Note on [XR Accessibility Requirement](#)
- Further reading:
  - [Extending WWW to support Platform Independent Virtual Reality](#), May 1994
  - [Open Standards for the Immersive Web](#), ERCIM News, April 2024
  - [Microsoft Teams enters the metaverse race with 3D avatars and immersive meetings](#)

\* The [omniverse](#) encompasses all universes, real and imagined, including the IoT and all metaverses.



# The Omniverse Awaits

Thanks for your attention