New W3C Standards for Speech and Multimodal Applications

Deborah A. Dahl
Principal, Conversational Technologies
Chair, W3C Multimodal Interaction Working Group

Voice Search Conference
March 10-12 2008 San Diego

Standards for Voice, Multimodal, and Mobile Applications

- Kinds of standards
- Overview of current and upcoming standards
- A closer look at: Multimodal Architecture, EMMA, SCXML
- Example: putting it all together – standards applied to Voice Search

Kinds of Standards

- **Architecture and Communication**
  - **What:** Describe how functions are allocated among hardware/software components and how they communicate
  - **Why:** Ensure interoperability of components
- **User Interface**
  - **What:** Guidelines for developing usable systems
  - **Why:** Ensure that user interfaces accommodate the perceptual, motor, and cognitive capabilities of human users and are consistent with social and cultural expectations
- **Application Definition**
  - **What:** Markup for defining applications
  - **Why:** Make applications easier to build
- **Certification**
  - **What:** Define a name or attribute
  - **Why:** Ensure that products with that attribute have known properties

Some Organizations Concerned with Standards

- World Wide Web Consortium Working Groups
  - Voice Browser
  - Multimodal Interaction
  - Ubiquitous Web Applications
  - Device Description
  - Web Accessibility
  - Mobile Web Best Practices
- IETF
- VoiceXML Forum
- OMA

Architecture and Communication

- **Multimodal Architecture**
- **Life Cycle Events – Communication among multimodal components**
- **EMMA – represents user input**
- InkML – describes stylus input
- DCCI – describes device context and interfaces
- Device Description – describes devices
- MRCP – messages to and from speech engines
- HTTP – basic message format for the Web

Application Definition

- **SCXML – Flow control for an application**
- VoiceXML – voice interaction
- HTML – GUI web pages
- SRGS/SISR – defines grammars and semantics for speech recognizers
- PLS – defines how words are pronounced
- CCXML – call control

User Interface

- Web Content Accessibility Guidelines 2.0
- Mobile Web Best Practices 1.0
- Common Sense Suggestions for Developing Multimodal User Interfaces

Certification

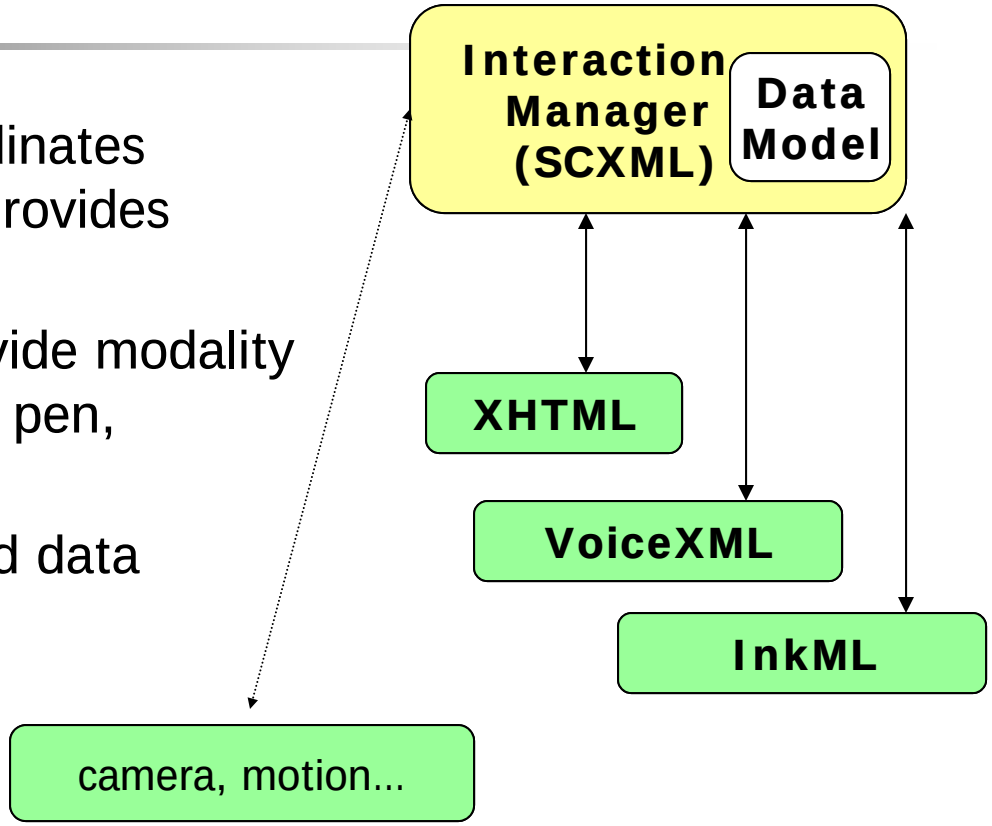
- Mobile Web Best Practices -- MobileOK
- VoiceXML Forum – platform certification and developer certification

Multimodal Architecture and Interfaces

- A loosely-coupled, event-based architecture for integrating multiple modalities into applications
- All communication is asynchronous, event-based
- Based on a set of standard life-cycle events
- Components can also expose other events as required
- Encapsulation protects component data
- Encapsulation enhances extensibility to new modalities
- Can be used outside a Web environment

MMI Architecture: Components

- *Interaction Manager*—coordinates modality components and provides application flow
- *Modality Components*—provide modality capabilities such as speech, pen, keyboard, mouse
- *Data Model*—handles shared data



MMI Architecture: Communication

- Life Cycle Events
- EMMA

Life Cycle Events

Event	From	To	Purpose
NewContextRequest	Modality	Runtime Framework	Request new context
NewContextResponse	Runtime Framework	Modality	Send new context id
Prepare	Runtime Framework	Modality	Pre-load markup
PrepareResponse	Modality	Runtime Framework	Acknowledge Prepare
Start	Runtime Framework	Modality	Run markup
StartResponse	Modality	Runtime Framework	Acknowledge Start
Done	Modality	Runtime Framework	Finished running
Cancel	Runtime Framework	Modality	Stop processing
CancelResponse	Modality	Runtime Framework	Acknowledge Cancel
Pause	Runtime Framework	Modality	Suspend processing
PauseResponse	Modality	Runtime Framework	Acknowledge Prepare
Resume	Runtime Framework	Modality	Resume processing
ResumeResponse	Modality	Runtime Framework	Acknowledge Resume
Data	either	either	Send data values
ClearContext	Runtime Framework	Modality	Deactivate context

EMMA (Extensible MultiModal Annotation)

- XML format
- represents results of processing user input
- includes annotations for information about the input (confidence, timestamp, tokens, language, etc.)
- can be used for input from speech, ink camera, keyboard...
- Contents of Data property of the Data or Done Life Cycle event

Application Definition: State Chart XML (SCXML)

- Extension of state transition systems for logic control
- Support multiple data input modalities, for example
  - VoiceXML for voice
  - HTML for GUI
  - InkML for handwriting
  - Extensible to other modalities such video as input, a GPS, kinesthetic sensor input on mobile devices, and other sensor devices
- States and transitions
  - State sends and receives messages from input/output modalities
  - Transitions to another state when conditions are satisfied

Putting Standards Together: A Voice Search application

- MMI Architecture
  - Interaction Manager (SCXML)
  - Modality Components (VoiceXML and HTML)
- Communication
  - MMI Life Cycle events + EMMA
- Application Definition
  - SCXML defines application flow
  - VoiceXML defines voice interaction
  - HTML defines GUI interaction

Startup

- User clicks on a web page to request a session
- SCXML Interaction Manager starts the application
- SCXML sends HTML page to web browser
- SCXML sends VoiceXML page to the voice browser, along with a general grammar

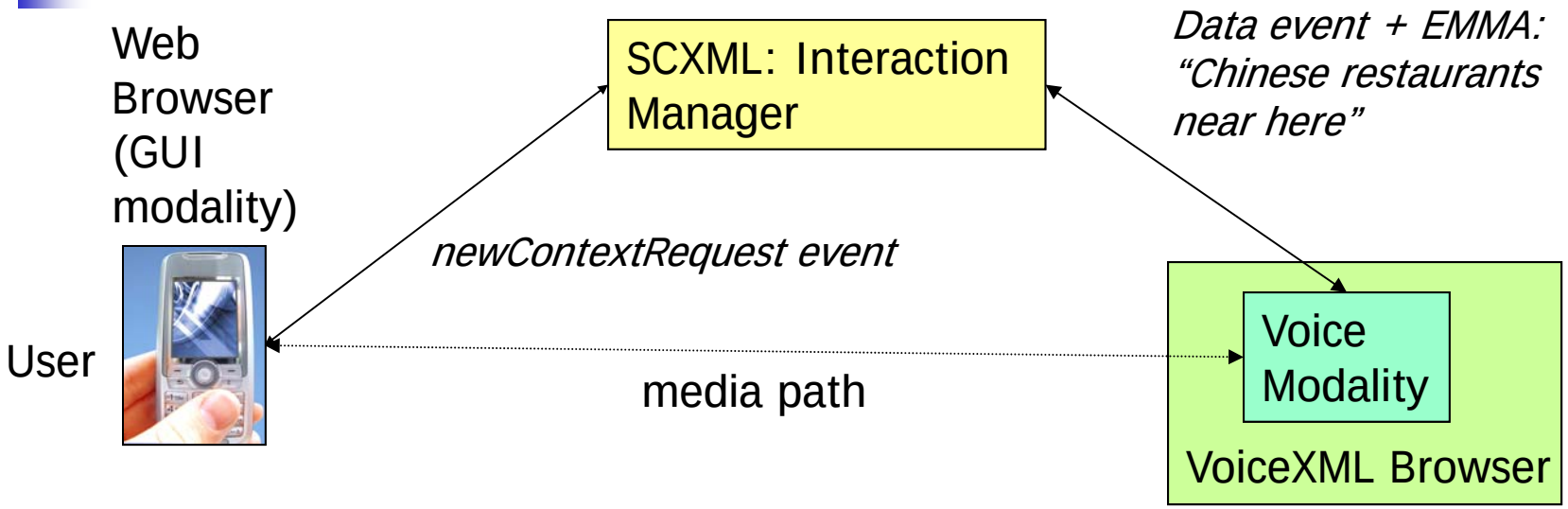
User's Perspective

"Chinese restaurants near here"

Basic Search Operation

- User speaks a search request
  - “show me Chinese restaurants near here”
- Voice browser converts request to EMMA and sends back to IM in a “data” life cycle event
- IM submits request to a search engine
- IM receives results as an HTML page
- IM sends HTML page with results to GUI browser
- User selects a result

Basic Search Interaction



More Complex Interactions: Integrating Voice and GUI

- “show me the first one”
- “this one” (click)
- “call the first one”
- “show me Cin Cin”
- “navigate to the first one”
- “how far is the Lucky Village?”
- “send this page to Sarah’s phone”

More Information

- W3C: www.w3.org
- IETF MRCP: www.ietf.org/internet-drafts/draft-ietf-speechsc-mrcpv2-15.txt
- Architectures and Communication
  - MMI Architecture: www.w3.org/TR/mmi-arch/
  - EMMA: www.w3.org/TR/emma/
  - Delivery Context: Client Interfaces: www.w3.org/TR/DPF/
  - Device Description: www.w3.org/2005/MWI/DDWG/
- Application Development
  - SCXML: www.w3.org/TR/voicexml21/
  - VoiceXML: www.w3.org/TR/voicexml21/
  - CCXML: www.w3.org/TR/voicexml21/
- User Interface
  - Commonsense Suggestions for Multimodal Applications: www.w3.org/TR/mmi-suggestions/
  - Web Accessibility Content Guidelines: www.w3.org/TR/WCAG20/
  - Mobile Web Best Practices: www.w3.org/Mobile/
- Certification
  - MobileOK: www.w3.org/Mobile/
  - VoiceXML certification: www.voicexmlforum.org