## **Canonical Document Type Guide for RAG System**

This document standardizes all structured prompts for internal content types processed in the SandS Media RAG system.

Each content type is organized according to five unified structural dimensions:

1. **Document Identity**
2. **Author Role Terminology**
3. **Parsed Fields & Metadata Descriptions**
4. **Content Qualities & Interaction Considerations**
5. **Contextual Notes & Differentiation**

### **✅ Chunking & Metadata Schema (Applies to All Content Types)**

To accommodate long documents such as articles, lessons, and workshops, the RAG system supports chunking of both text and optionally slidetext. This section outlines the standard schema for managing chunked documents.

**Note on Terminology**:  
Throughout this guide, a “document” refers to a single logical unit such as an article, lesson, workshop day, or recorded session. However, for purposes of RAG retrieval and vector indexing, these documents may be **chunked into smaller parts**. All metadata fields listed below (including title, author, parentName, etc.) are inherited by every chunk. Content fields like text and slidetext may be divided across multiple chunks.

Therefore, whenever a content type description says “this document represents...”, it should be understood as **the logical full document**, not an individual chunk.

#### **1. Retained Standard Metadata (Shared Across All Chunks)**

| **Field** | **Description** |
| --- | --- |
| contentType | One of: READ, TUTORIAL, FSLE, RHEINGOLD, CAMP, FLEX\_CAMP |
| documentId | Unique ID for the original document (shared by all its chunks) |
| title | Title of the full document |
| abstract | Abstract of the full document |
| parentName | Name of the enclosing magazine issue/tutorial/online live event/conference/seminar |
| parentDescription | Static backend-provided brand or series descriptor |
| author | Name of the author/speaker/trainer |
| access | Either granted or restricted — based on the user’s subscription status or whether the add on content has been purchased (refer to User Context Field Guide Document for more details) |
| language | Language of the original document |
| date | Start date of document or recording availability |

#### **2. Chunkable Text and Slide Fields**

| Field | Description |
| --- | --- |
| chunkSource | Indicates the content source of the chunk: either "text" or "slidetext" |
| text | Text content for chunkSource: text (vectorized) |
| slidetext | Slide content for chunkSource: slidetext (vectorized, typically not chunked unless lengthy) |

#### **3. Chunk Positioning Metadata**

| Field | Description |
| --- | --- |
| part\_number | Index of the chunk within its source (e.g., 1) |
| total\_parts | Total number of chunks for that source (e.g., 3) |

* part\_number and total\_parts refer to the set of chunks that share the same chunkSource.
* If the source is unchunked: part\_number = 1 and total\_parts = 1

#### **4. Retrieval & Prompt Implications**

* **Document ID Anchoring**: Chunks with the same documentId must be grouped for contextual understanding.
* **Chunk-aware System Prompting**: LLMs must infer document structure and reconstruct partial information where needed.
* **Caveat for Retrieval**: Only one chunk may be surfaced at a time. Prompt logic should prevent overgeneralization from partial data.
* **Temporal Consideration**: For TUTORIAL, FSLE, RHEINGOLD, and FLEX\_CAMP, text and slidetext may initially only contain title and description of the document and become richer, by way of adding audio transcript, only after the date in date field.

### **✅ READ (Article)**

**1. Document Identity**

* Document type: READ
* This is a professionally edited *article* (EN) / *Artikel* (DE), published in a digital magazine format.
* Distributed via the entwickler.de (DE) or devm.io (EN) platforms.

**2. Author Role Terminology**

* Refer to the author as “Author” (EN) / “Autor” (DE)

**3. Parsed Fields & Metadata Descriptions**

* contentType: always READ
* date: Date of publication
* title: Title of the article
* abstract: Short abstract or summary
* parentName: Magazine brand and issue (e.g., "Java Magazin 01.2024")
* parentDescription: Standardized backend text describing the brand
* text: Main body of the article. This field may contain **only one part** of the full article due to document chunking.
* author: Author’s name
* access: granted for user accessTier value basic or higher (refer to User Context Field Guide Document for more details) unless public or promotional article
* language: Language of the article
* documentId: Unique identifier of the full document (shared across all chunks)
* slidetext: *Always empty for READ content types*

**4. Content Qualities & Interaction Considerations**

* Topics: Solutions to development problems, tool updates, frameworks
* Tone: Neutral, objective, professional
* Structure:
  + Title
  + Abstract (Intro)
  + Segmented main body
  + Conclusion
  + Code listings with English comments
  + Illustrations and diagrams
  + If present, author bio appears at end
* Current magazine brands sorted by their respective frequency and publication language:
  + Monthly (12/year) German: Java Magazin, Windows Developer, Entwickler Magazin
  + Bi-monthly (6/year) German: DevOpsCon Magazin, MLcon Magazin, iJS Magazin, PHP Magazin
  + Bi-monthly (6/year) English: DevOpsCon Magazine, MLcon Magazine, iJS Magazine, PHP Magazine

**5. Contextual Notes & Differentiation**

* Not episodic; each article stands alone
* Not instructional like tutorials/workshops
* Not bound to a learning path
* Do not reference this as a brand-neutral source — brand affinity matters in community display contexts
* Ideal for readers seeking standalone insight or professional updates

### **✅ TUTORIAL (Video Lesson)**

**1. Document Identity**

* **Document type:** TUTORIAL
* This document represents one lesson within a structured, multi-lesson video tutorial.
* Distributed via the entwickler.de (DE) or devm.io (EN) platforms.

**2. Author Role Terminology**

* Refer to the author as **"Speaker"** (EN/DE)

**3. Parsed Fields & Metadata Descriptions**

* contentType: always TUTORIAL
* date: Date this lesson was made available (recordings and transcript become available only after this date)
* title: Title of the video lesson
* abstract: Summary of the lesson
* parentName: Tutorial series name
* parentDescription: Description of the tutorial series
* text: Transcript of the lesson (automatically generated; may contain linguistic errors). This field may contain **only one part** of the recording transcript due to document chunking.
* slidetext: May contain uploaded slides (chunked if needed; empty if not applicable)
* author: Speaker’s name
* access: granted for user accessTier value fullstack or elevate (refer to User Context Field Guide Document for more details)
* language: Language of the spoken tutorial
* documentId: Unique identifier of the full document (shared across all chunks)

**4. Content Qualities & Interaction Considerations**

* Each document is a **lesson** within a full **tutorial** (typically 3–6 lessons per tutorial)
* Lessons are subdivided into **chapters** (3–20 minutes each)
* **Visual elements** (slides, code demonstrations) are essential to understanding but not present in transcript
* Content ranges from **introductory to advanced**; assumes baseline developer knowledge
* **Transcript and slidetext** become available only *after the date field*
* Medium awareness required: treat content as one part of a **multi-modal, episodic** learning experience

**5. Contextual Notes & Differentiation**

* Tutorials are **episodic**: lessons build upon each other; never treat as standalone
* Branded similarly to conference brands (DevOpsCon, JAX, iJS, etc.) — though not included in metadata yet
* Strongly contrasted with:
  + **READ**: not an editorial article
  + **FSLE**: not a one-off live event
  + **RHEINGOLD**: not a time-limited conference session
  + **CAMP/FLEX\_CAMP**: not interactive, cohort-based training
* Emphasize guided progression, clarity of instruction, and educational depth

### 

### **✅ FSLE (Live Event Session)**

**1. Document Identity**

* **Document type:** FSLE
* FSLE stands for Fullstack Live Event. This document represents a single recorded session from a full-day **Live Event** held online in real time.
* Events may consist of multiple sessions, typically organized around a central theme.

**2. Author Role Terminology**

* Refer to the author as **"Speaker"** (EN/DE)

**3. Parsed Fields & Metadata Descriptions**

* contentType: always FSLE
* date: Date of the live event (recordings and transcript become available only after this date)
* title: Title of the session
* abstract: Summary of the session’s topic
* parentName: Name of the full-day Live Event under which the session was delivered
* parentDescription: Description of the Live Event (static backend value)
* text: initially same content as title plus abstract, later transcript of the session is added (automatically generated; may contain linguistic errors). This field may contain **only one part** of the recording transcript due to document chunking.
* slidetext: May contain uploaded slides (chunked if needed; empty if not applicable)
* author: Speaker’s name
* access: granted for user accessTier value fullstack or elevate (refer to User Context Field Guide Document for more details) unless public or promotional online live event
* language: Language of the session
* documentId: Unique identifier of the full document (shared across all chunks)

**4. Content Qualities & Interaction Considerations**

* Single session recordings, typically **45–90 minutes**
* Sessions are part of themed full-day online events
* May include real-time **Q&A** or panel interactions — these are *not* reflected in the transcript
* **Slides and transcript** become available only *after the date field*
* Events vary in style: lectures, interactive presentations, practical walkthroughs

**5. Contextual Notes & Differentiation**

* FSLE sessions are **standalone recordings**, not part of a structured tutorial path
* Unlike TUTORIAL: not episodic or progressive
* Unlike RHEINGOLD: not part of a multi-day conference; shorter and less formal
* Unlike CAMP or FLEX\_CAMP: not hands-on, no certificates, and not cohort-based
* Best understood as **live insights**, condensed knowledge from experienced professionals

### 

### **✅ RHEINGOLD (Conference Content)**

**1. Document Identity**

* **Document type:** RHEINGOLD
* This is a **recorded session transcript** from a software professional conference or summit. Attendance mode is online or in person.
* It belongs to one of two subtypes:
  + **Regular Conferences** (multi-track, multi-day events with sessions, workshops, and keynotes)
  + **Training Conferences (Summits)** (intensive, workshop-heavy training events)

**2. Author Role Terminology**

* Regular Conference: Refer to the author as **"Speaker"** (EN/DE)
* Training Conference (Summit): Refer to the author as **"Trainer"** (EN/DE)

**3. Parsed Fields & Metadata Descriptions**

* contentType: always RHEINGOLD
* date: Date of the recorded session
* title: Title of the session, keynote, or workshop
* abstract: Summary of the session’s topic
* parentName: Conference or Summit brand + location/year (e.g., "JAX Munich 2024")
* parentDescription: Static backend description of the event
* text: initially same content as title plus abstract later transcript of the recording (auto-generated,may contain linguistic errors, available **after** date) is added. This field may contain **only one part** of the full article or recording transcript due to document chunking.
* slidetext: May contain uploaded speaker slides (added post-event)
* author: Name of the speaker or trainer
* access: granted for users who specifically purchased attendance of the event (refer to User Context Field Guide Document for more details)
* language: Language of the session
* documentId: Unique identifier of the full document (shared across all chunks)

**4. Content Qualities & Interaction Considerations**

* Formats include:
  + **Sessions** (30–60 minutes)
  + **Workshops** (half-day)
  + **Keynotes** (30–45 minutes)
* Multi-track event: many sessions may occur in parallel
* Slide and code references often essential for full comprehension
* Transcript and slides become available only **after** the session date
* The session type (keynote/workshop/etc.) is **not explicitly structured** in metadata, but may be inferred from title or abstract

**5. Contextual Notes & Differentiation**

* Conference brands include:
  + **Regular Conferences:** JAX, BASTA!, DevOpsCon, iJS, MLcon, IPC, APIcon, Serverless Architecture Conference, webinale
  + **Training Summits:** Software Architecture Summit, JavaScript Days, MAD-Summit
* Not a general access product: RHEINGOLD content is a **premium add-on**, not included in the fullstack membership
* If the brand is a Summit, the content is likely a **workshop** unless explicitly noted otherwise
* These sessions are **event-based** knowledge snapshots — not serialized tutorials
* When referencing, highlight that this is **premium, in-person event content**, distinct from regular platform access

### 

### **✅ CAMP (In-Person Workshop)**

**1. Document Identity**

* **Document type:** CAMP
* This document represents a **full-day, in-person workshop**, part of a multi-day training program called a **Camp**.

**2. Author Role Terminology**

* Refer to the author as **"Trainer"** (EN/DE)

**3. Parsed Fields & Metadata Descriptions**

* contentType: always CAMP
* date: Start date of the Camp event
* title: Title of the specific workshop day (e.g., “Domain-Driven Design – Day 1”)
* abstract: Summary of the workshop day’s focus
* parentName: Camp name with location and year (e.g., "DevOps Camp Berlin 2024")
* parentDescription: Static backend description of the Camp series
* text: *Empty* — workshops are **not recorded**
* slidetext: May contain uploaded slides (chunked if necessary; otherwise single block)
* author: Trainer’s name
* access: granted for users who specifically purchased attendance of the event (refer to User Context Field Guide Document for more details)
* language: Language of instruction
* documentId: Unique identifier of the full document (shared across all chunks)

**4. Content Qualities & Interaction Considerations**

* Duration: 2–5 days (each document = 1 full-day workshop)
* Intensive format with:
  + Practical exercises
  + Real-world case studies
  + Trainer-led walkthroughs
* Delivered in-person, online or hybrid (typically hosted in modern training facilities)
* Additional benefits:
  + Certificates of participation (Teilnahmezertifikat)
  + Some include formal certifications (e.g., certified professional software architect **CPSA** by awarding body **ISAQB**)
  + Networking and evening events included

**5. Contextual Notes & Differentiation**

* Unlike READ, FSLE, or TUTORIAL: **not** publicly recorded or editorial
* Unlike FLEX\_CAMP: **not asynchronous** or self-paced
* Not a general access product: CAMP content is a **premium add-on**, not included in the fullstack membership
* Use a **marketing-forward tone** to emphasize exclusivity, expert guidance, and high-value outcomes

### 

### **✅ FLEX\_CAMP (Self-Paced Online Workshop)**

**1. Document Identity**

* **Document type:** FLEX\_CAMP
* This document represents a **self-paced video lesson or chapter**, released as part of a structured **Flex Camp** asynchronous training series. This content is not available through platform membership.

**2. Author Role Terminology**

* Refer to the author as **"Trainer"** (EN/DE)

**3. Parsed Fields & Metadata Descriptions**

* contentType: always FLEX\_CAMP
* date: Date when this session/chapter becomes available to learners
* title: Title of the lesson within the Flex Camp
* abstract: Short summary of the specific lesson
* parentName: Flex Camp name and launch reference (e.g., "MLcon Flex Camp 06.2024")
* parentDescription: Static backend-provided description of the Flex Camp series
* text: initially same content as title plus abstract, later transcript of the lesson is added (auto-generated, available **after** date). This field may contain **only one part** of the recording transcript due to document chunking.
* slidetext: May include slides if provided (typically unchunked)
* author: Trainer’s name
* access: granted for users who specifically purchased attendance of the event (refer to User Context Field Guide Document for more details)
* language: Language of instruction
* documentId: Unique ID for the full lesson

**4. Content Qualities & Interaction Considerations**

* Fully self-paced structure:
  + New lessons unlocked weekly after initial launch
  + Learners control pace and revisit material
* Learning aids include:
  + Bookmarking and progress tracking
  + Q&A support channels (responses within 24h)
  + Periodic **live sessions** for clarification
* Video, transcript, and slidetext become available **only after** release date
* Designed for solo learners or upskilling teams

**5. Contextual Notes & Differentiation**

* Not a general access product: FLEX\_CAMP content is a **premium add-on**, not included in the fullstack membership
* Unlike TUTORIAL: not episodic in linear series, but asynchronous access
* Unlike CAMP: no live or in-person interaction
* Unlike RHEINGOLD or FSLE: not recorded from live event
* Clearly emphasize its **modern, flexible format**, ideal for busy professionals or distributed teams