

# Towards Supporting Context-oriented Information Retrieval in a Scientific-Archive based Information Lifecycle

Felix Engel, Claus-Peter Klas, Holger Brocks, Alfred Kranstedt, Gerald Jäschke, Matthias Hemmje

# Agenda

- SHAMAN
- Objectives
- Context Model
- Context-oriented Information Retrieval
- Conclusion

## SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg)

- Co-funded by the European Commission under the seventh RTD Framework Programme
- **Aim:** next generation digital preservation framework
- Three **Integration & Demonstrator Subprojects** (ISPs)
  - **ISP1:** *Document Production, Archival, Access and Reuse in the Context of Memory Institutions for Scientific and Governmental Collections*
  - **ISP2:** Simple and Connected Object Production, Archival and Reuse in the Industrial Design and Engineering Domain
  - **ISP3:** eScience Data-Acquisition and Harmonisation Test bed
- Base on ISP1 with focus on the domain of **Scientific Publishing**

## Objectives

- **Foster reuse of preserved resources**
  - By providing a **proper description** of preserved digital resources
    - → **Context model**
  - Used as basis for supporting **retrieval capabilities** on archive access

## SHAMAN: Context Model

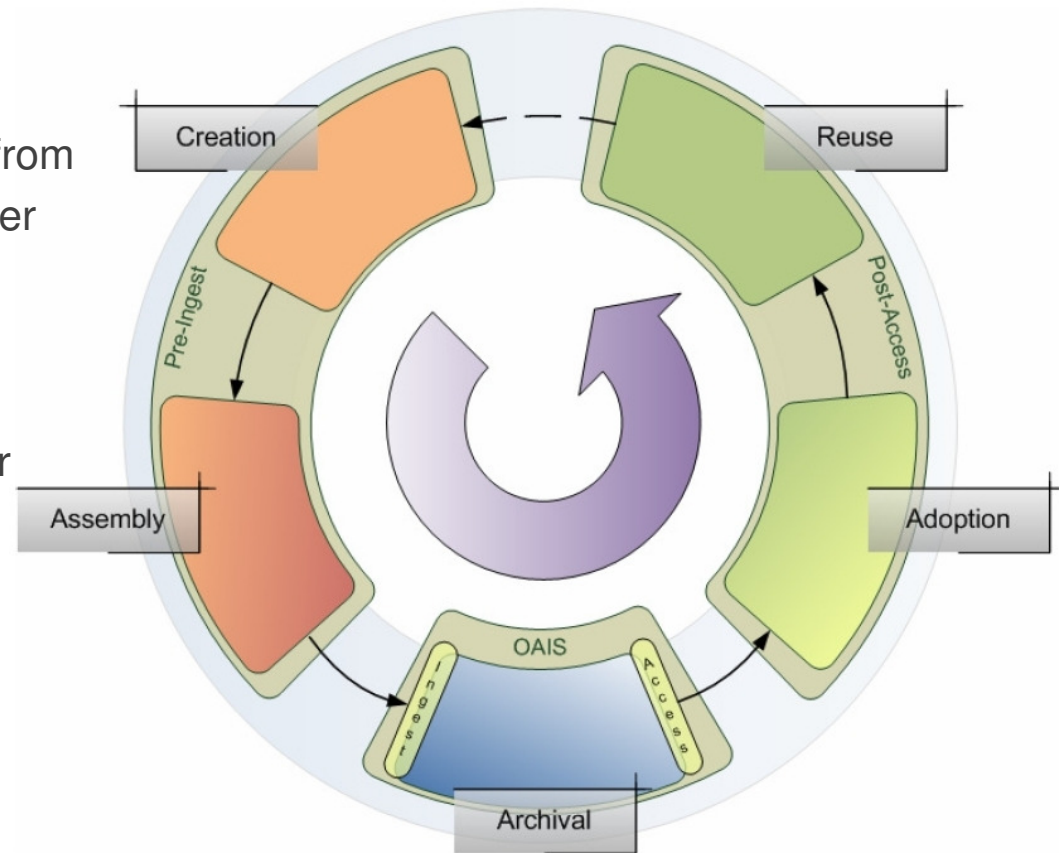
- **Problem of change**
  - Digital resource nature and usage settings could **change over time**
  - Consumers **cannot refer back** to the creators!
  - Reuse depends on **proper description**
    - → Should be provided through the preservation system
- **Support by modelling context**
  - **SHAMAN context model**
    - Provides an infrastructure-independent representation of the attributes associated with and (implied) relations between digital objects
    - Enables capturing and representing of a wide range of additional data

## Context

- Multitude of **distinct notions across domains**
  - SHAMAN notion accords to: *interrelated conditions in which something exists or occurs* [2]
  - → **Intersection** of most context definitions
- **Implies a complex structure** of context
  - Share concepts with other objects
    - **Process environment** (in which they are created)
    - **Associated actors, resources and information objects**
    - **Preservation environment** in which they are stored
  - → Context representation: **not a separate entity**
- Need for context capturing and representation through **all life phases**
  - → Context data are **lost** if they are **not captured in time**
  - → **Information Life Cycle Model**

# Information Life Cycle Model

- Today, archives derive metadata from the digital object preserved together with some metadata
- Context data of digital objects evolve in **different phases** of their existence
- Phases are Modeled in the **Information Life Cycle Model**



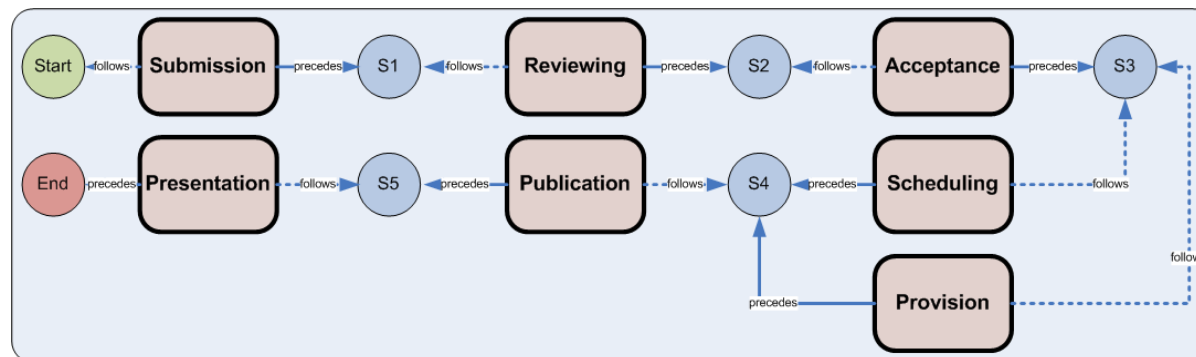
# Use Cases for Context-oriented Information Retrieval

## Organizing the review process for a congress

- **Peer reviewing** process
  - Assign conference contribution to reviewer (with appropriate expertise)
- Relevant **selection criteria** in order to organize the conference reviewer team
  - **Previous review experience**
    - How often (number of review activities)
    - Where (which conferences)
  - **Domain knowledge**
    - publications that the candidate has previously reviewed
    - the candidate has authored
    - the candidate has referenced in his own work or that reference the candidate's own publications

## Domain of Scientific publishing

- **Scientific publications** nowadays expected origin **born digital**
- Presented and discussed at **conferences**, preserved in **archives**



Simplified Scientific Publishing Process – Control-Flow

## Context Data in the Domain of Scientific publishing

- **Established processes** in course of a conference
  - ... incurred data
    - **Writing** (amongst others)
      - Abstracts, papers, posters, presentations
    - **contributions** are
      - Reviewed, accepted, indexed
    - **Scheduling**
      - Speakers get invited, a scientific program is set-up, categorized and linked thematically
    - **Publication** and **presentation**
      - Authors, presenters, conference participants, events, sessions, talks, and topics
      - Structured text information (conference contributions), abstracts, tables and figures

## Concepts of context

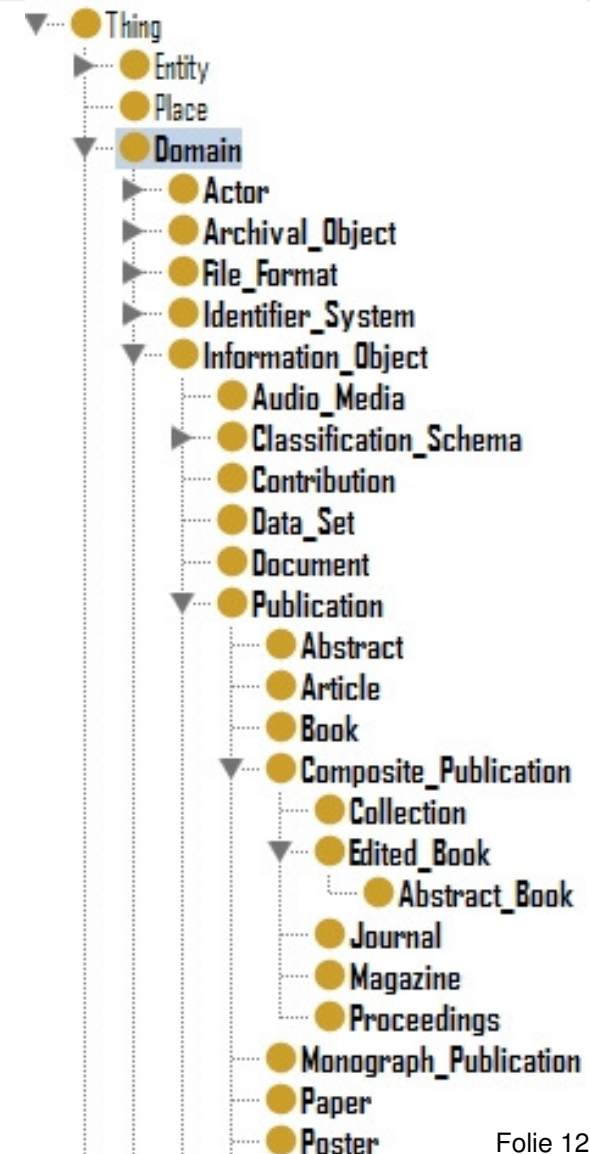
Three **distinct concepts** strongly involved in defining context:

(→ With reference to the domain of scientific publishing, by means of a conference)

- **Domain:** concepts specific to the domain and their relations. For instance  
→ *Abstract, Abstract Book, Presentation or Supplement, conference topics or review criteria*
- **Enterprise:** the structural layout of an organizational environment. For instance:  
→ *Affiliation, Persons or Roles*
- **Process:** the processes and their associated activities, including  
→ Information about their implementations (service invocations): *Submission, Indexing or Reviewing*

## Context Representation

- **Ontology** is used for serialization, representation and organization
- Ontology in short
  - **Formal model** of a specific domain
  - Represent *concepts* and their *relations* to one another



## Context-oriented Information Retrieval

- SHAMAN context model provides numerous **additional information**
  - Context data
  - Relations between context data (structured access)
- Retrieval could be **extended** in multiple ways
  - Creation of an **additional full-text** index, containing the indexed context data
  - Retrieval approaches based on **ontological relations**
  - Ontology based **visualization**
- Retrieval on base of ontology
  - Moving by means of relations
  - Logic based reasoning

# Context-oriented Information Retrieval

Exemplary retrieval approaches on **bases of ontology** [3]

- **Generalize: move upwards**
  - Move from a conference presentation to the conference session
  - E.g. **move from presentation** *Ventral striatum and impulsivity: fMRI in pathological gambling* **to session** *Neuroimaging the impulsivity in psychiatry*
- **Specify: move downwards**
  - Move from a conference session to a specific publication
  - E.g. **move from session** *Neuroimaging the impulsivity in psychiatry* **to presentation** *Ventral striatum and impulsivity: fMRI in pathological gambling*
- **Move sideways**
  - Move from conference session X to session Y
  - E.g. **move from session** *Neuroimaging* **to session** *Genetics*

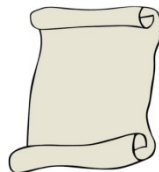
## Context-oriented Information Retrieval

- Use those retrieval approaches as **basis for complex retrieval processes**
  - Link single retrieval steps
- Retrieval processes for instance
  - **Start** with full-text retrieval
    - If an appropriate result is found → **refine** with ontology based retrieval approaches
  - **Start** with an appropriate ontology concept
    - → **Refine** with ontology based retrieval approaches

→ **Result: archived objects or set of context data**

**Poster:**

*Tobacco smoking and motivation  
to quit in psychiatric patients*



**Poster presentations:**

*Addictive Disorder  
Affective Disorders (Bipolar)  
Affective Disorders (Unipolar)  
Anxiety  
Childhood Adolescent Disorders*

....

## Conclusion

Context data are valuable for the **reuse** of preserved resources

Context data of a digital resource should be captured in all **life phases of the resource** (Information Life Cycle Model)

Context data are a potentially invaluable source in order to **extended retrieval in archive access**

Thanks for your attention !

## References

- [1] CCSDS. Reference Model for an Open Archival Information System (OAIS). Blue Book 1, Consultative Committee for Space Data Systems, January 2002. Recommendation for Space Data Systems Standards, adopted as ISO 14721:2003.
- [2] Merriam-Webster Online Dictionary. context; cited 30.04.2009. "ONLINE"  
<http://www.merriam-webster.com/dictionary/context>.
- [3] Elena García and Miguel-Ángel Sicilia. User Interface Tactics in Ontology-Based Information Seeking. PsychNology Journal, 2003 Volume 1, Number 3, 242 - 255

# Use Cases for Context-oriented Information Retrieval

## Retrieval tasks on a specific topic

- Use the **full-text index** in order to find a first relevant publication as **starting point**
  - For instance a first relevant publication was found, published at the *9th World Congress of Biological Psychiatry*
- Finding **more relevant publications of the same topic**
  - Managed by **moving down the ontology congress hierarchy**
  - For instance
    - The talk of the relevant publication was held in the **symposium *Neuroimaging the impulsivity in psychiatry***
    - Additional information through examination of further publications presented within the **same session**
    - ... or within **other sessions** arranged under the conference topic *Neuroimaging*

**Creation:** new information comes into existence

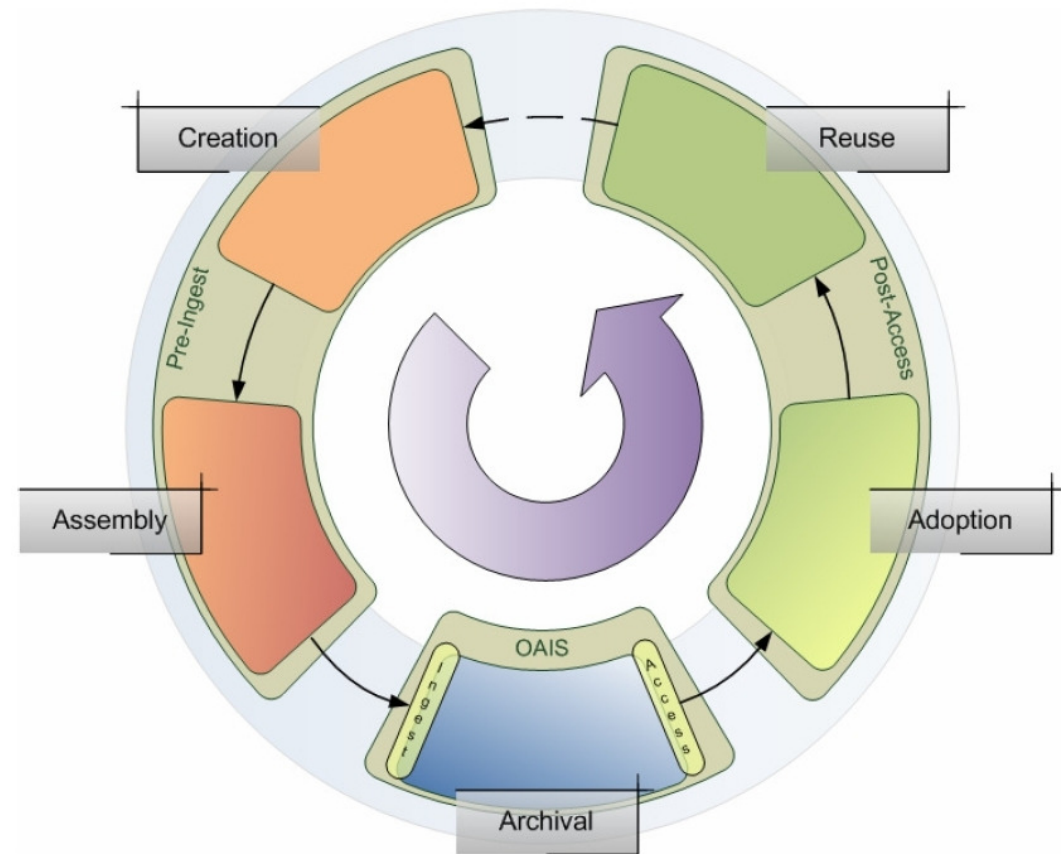
**Assembly:** appraisal of objects relevant for archival and all processing and enrichment for compiling the complete information set to be sent into the future

**Archival:** life-time inside the archive

**Adoption:** processes by which accessed archival packages are unpacked, examined, adapted, transformed, integrated and displayed to be usable and understandable for the consumer

**Reuse:** exploitation of information by the consumer. Reuse of Digital Objects can lead to the Creation of other, novel Digital Objects. Reuse also may instigate the addition or updating of metadata about the Digital Object held in the archive

## Information Life Cycle Model



## Context Data in the Domain of Scientific publishing

**Context data** of scientific publications by means of a conference, provided by

- Abstract Book (print or net publication)
- The *conference web site*, interactive and structured access to the abstracts along
  - Date, time, type of presentation, topic, presenter
- Consecutive editions of one conference
  - Additional metadata
  - → Multitude of relationships