

# Preserving and Annotating Publishers' Data

Presenters:

Peter Cornwell (ENS-Lyon/Data Futures)

José Gonzalez (CERN)

Tom Lamberty (Merve Verlag)

FREE SEMINAR

click on link below to register



16th of October 2020

15:00-16:00 CEST





## Preserving and Annotating Publishers' Data

- **Peter Cornwell** co-chair RDA PTTP-IG, ENS-Lyon, IMCC Westminster
- **José Gonzalez** Head of repository technologies, CERN
- **Tom Lamberty** Managing Director, Merve Verlag
- email contact form: [www.data-futures.org/pages/contact](http://www.data-futures.org/pages/contact)



## Preserving and Annotating Publishers' Data

- **open reading online** (albeit with access control options) using presentation-capable IIF service supporting multiple standards-based readers
  - [catalogue.merve.hasdai.org](http://catalogue.merve.hasdai.org)
- **print-on-demand** of volumes which are no longer in print, plus custom reader print service—whether current catalog or back-list
- **support for researchers** via standards-based interfaces to library systems, creation, maintenance and publishing of annotations and full corpus search



## Preserving and Annotating Publishers' Data

- **preserve the publisher:** *back-catalog* (born-analog 1970-1990's) plus *current books* (born-digital 1990's-today) redelivered as agnostic HTML, multiple styled outputs as well as book metadata, using a trusted corpus-specific repository (Invenio—same tech as Zenodo)
  - preserve historic scans, PDFs, raw OCR and original digital sources—*including markup*
  - maintain and preserve new digital resource: agnostic HTML, styled print-ready outputs incl. e.g. custom readers
- sustainable archive of publisher business administration documents, including agreements and licenses, correspondence with agents such as authors, editors and translators as well as distribution data like e.g. reviews

## transform existing books

- workflow-1: accession of *back-catalog* scans to produce page impressions and manage OCR data using MongoDB-based *freizo* platform
  - fix recognition errors, remove end-of-line hyphenation and make flexible linked foot/end note definitions, nested section styles and emphasis
  - generate HTML and IIF service for reading and research applications
- workflow-2 : redelivery of digitally originated *current-catalog* books
  - page impressions rendered from digital sources
- both workflows produce common outputs—ASCII, un-styled HTML & page imagery

## transform existing books

- produce new virtual catalog of technology-agnostic texts
  - automated assistance to correct recognition errors, transform hyphenation and edit and maintain historic books
  - create style-able HTML references and footnotes, electronic ToCs
- page impression, initial HTML and editor, plus heat-mapped OCR delivered side-by-side, together with page continuation and footnote/reference controls

Müller: Most people coming from West to East Berlin keep comparing them on a horizontal level. It doesn't work that way. The problem is the misery of comparing. You just can't compare things.

Lotringer: So East Berlin is something of its own?

Müller: I was very impressed by the remark of a young man who was writing an essay on my work. What was most interesting for him, he said, was this problem of another time. He remembered that he never quite understood why the German Wehrmacht didn't succeed in entering Moscow during the Second World War. They just stood there. They couldn't go further. He didn't believe in military or strategic reasons. He didn't believe in geographic reasons. He didn't believe in ideological reasons. There simply was a time-wall. They



## high-availability IIF service for online access

- IIF book manifests generated by *freizo* with tables of contents
  - support multiple reader and research applications using both Image and Presentation protocols ... open, standards-based APIs
  - 131 institutions are now members of growing IIF consortium: existing and future IIF-compatible applications include image viewers, and research instruments; much larger numbers of applications use IIF image fragments
- multiple IIF services generated automatically from conservation-quality master imagery provide parallel performance and automatic fail-over reliability for key resources — currently evaluated in bio-diversity community
- *hasdai*-supported Merve IIF service supports institutional guarantee to manage long-term costs through shared maintenance of multiple near-line copies

- ☀ Front matter
- ☀ Erste Vorlesung [23.04.2002]
- ☀ 0 Einführung
  - ▶ 1 Griechenland
  - ▶ 2 Philosophie
- ☀ Zweite Vorlesung [30.04.2002]
  - ▶ Dritte Vorlesung [07.05.2002]
- ☀ Vierte Vorlesung [14.05.2002]
  - ▶ 3 Mittelalter
- ☀ Fünfte Vorlesung [21.05.2002]
  - ▶ 4 Eine literaturgeschichtliche Revolution



Vorlesungen müssen  
genau das sagen  
was in keinem Buch  
umgekehrt

381

Merve

Philosophien der Literatur

Philosophien der Literatur

- ▶ 5 Kants Nachfolger
- Achte Vorlesung [11.06.2002]
- ▶ 6 Friedrich Wilhelm Georg Hegel
- ▶ **Neunte Vorlesung [18.06.2002]**
- Zehnte Vorlesung [02.07.2002]
- ▶ 7 Literaturphilosophie als Antiphilologie
- ▶ Elfte Vorlesung [09.07.2002]
- Zwölfte Vorlesung [16.07.2002]
- ▶ 8 Nach 1945
- 8.3.2 Roman Jakobson und Claude Lévi-Strauss
- 8.3.3 Jacques Lacan
- Editorische Notiz
- Literatur

7.3.1 DAS DASEIN IN SEINER ALLTÄGLICHKEIT .....	247
7.3.2 DER URSPRUNG DES KUNSTWERKS .....	254
7.3.3 DICHTUNG .....	259
<b>ZWÖLFTE VORLESUNG [16.07.2002] .....</b>	<b>263</b>
8 NACH 1945 .....	264
8.1 JEAN-PAUL SARTRE .....	264
8.2 ROLAND BARTHES ALS ÜBERGANG ZUM STRUKTURALISMUS .....	269
8.3 STRUKTURALISTEN .....	270
8.3.1 FERDINAND DE SAUSSURE .....	271
8.3.2 ROMAN JAKOBSON UND CLAUDE LEVI-STRAUSS .....	275
8.3.3 JACQUES LACAN .....	281
<b>EDITORISCHE NOTIZ .....</b>	<b>286</b>

Page 3 Page 4 Page 5 Page 6 Page 7 Page 8 Page 9 Page 10 Page 11

- Mirador (Harvard-Stanford) IIF viewer navigation using multi-level table of contents



## internationally-supported data preservation infrastructure

- Invenio-3.3 repository generated by *freizo* provides public access and large-scale search and networking facilities developed by the global community
  - mature services for connection with research community standards
  - comprehensive and tailorable search tools: Elasticsearch
- 2021 InvenioRDM release enables maintenance of catalog & archive by publisher
  - Invenio in-band maintenance of book sources and metadata
  - uploading of publisher business administration documents
  - access controls for selective reader copyright protection and work-in-progress

## Findable, Accessible, Interoperable, Reusable

- concurrent full corpus searches via independent MongoDB engine using ASCII text generated from book redeliveries
- serializer infrastructure (currently provided by freizo; built in to InvenioRDM) for generation of LoC metadata equivalents such as MODS and MARC and, significantly—IIIF manifests
- OAI Protocol for Metadata Harvesting enables external organizations to discover Merve records

## research interfaces & scholarly annotation

- sustainable reference IIF service for current and back book catalogs, with full presentation protocol layer, supporting creation and maintenance of annotation linked to contributors' ORCIDs
  - enables multiple existing and future workflows to link annotations to the original page interactively
  - preservation of canonical WADM annotation records using repositories such as Zenodo promotes collaboration and publishing and, significantly, serialization of multiple viewer software-specific annotation representation 'flavors', such as OADM dialects, from WADM

**Annotations**

- Die Kunst, Listen zu erstellen (1)
- Weiblichkeit in der Schrift (1)
- Rotwelsch (12)
- Geschwindigkeit und Politik (1)
- Kleine Schriften (2)
- Das Geschlecht, das nicht eins ist (3)

**Rotwelsch, 1982**  
download annotations as WADM

p11, a0 (Merve Digital Edition)  
**You really go through a time wall.**  
(vgl. Ernst Jünger: An der Zeitmauer (1959))

p13, a0 (Merve Digital Edition)  
**It's an artificial freedom, an artificial free space for ideology, for the arts and for literature.**  
(The term 'artificial' would be worthwhile a separate intervention)

p49, a0 (Merve Digital Edition)  
(He was the first to see/say this)

p107, a0 (Merve Digital Edition)  
(Das ist mal eine andere 'Traumdeutung')

- builds on existing support for WADM annotation: <https://de.merve.de> in 2018



## what's planned in 2021

- single-click creation and versioning of WADM annotation collections in Zenodo, with automated DOI generation
- search API plus print-on-demand fulfillment from Invenio catalog
- Kittler Special Edition workflows producing Collected Works print and open reading volumes plus Invenio repositories of non-printable media
- Fotomuseum Winterthur Still Searching Blog Special Edition
  - print and open reading volumes plus Invenio repositories
- preservation of Merve's historical documents, including agreements and licences and correspondence with agents such as authors, editors and translators