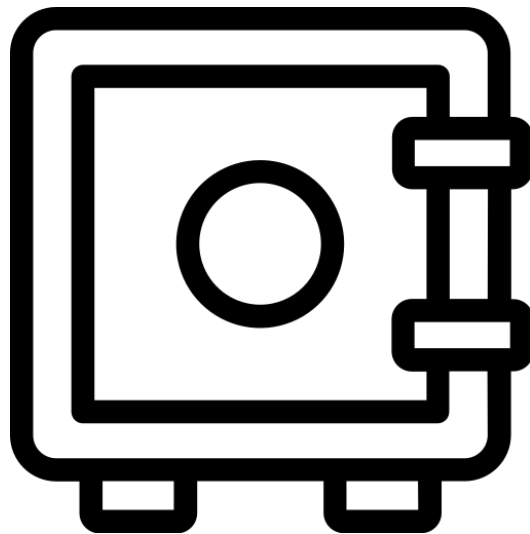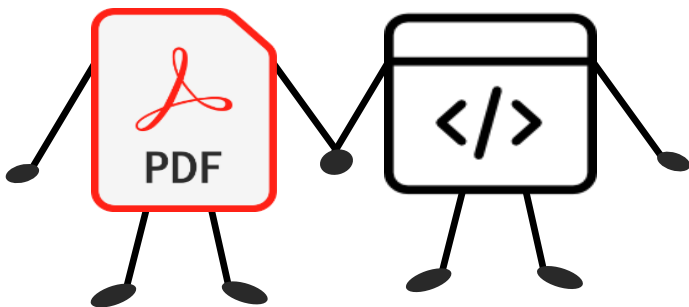# Keep it save and (not so) simple - How we archive our licensed content at the ZBW

Workshop Perpetual Access and Hosting of Electronic and Paper Resources
Luisa Kramer-Ibig, Kirsten Jeude
21.11.2023

Die ZBW ist Mitglied der Leibniz-Gemeinschaft.

Systematic archiving of e-resources (SAVE)

Agenda

1. SAVE: Motivation, subject, objectives, participants

2. Preliminary work: IT working environment, license management

3. Metadata management workflow: Acquisition, checking for completeness, review of PDFs, converting metadata

4. Archiving at the repository *Digital Archive* and digital long-term archiving with *Rosetta* and its international proof of evidence at *Keepers*

5. Lessons learned

Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

The project SAVE …
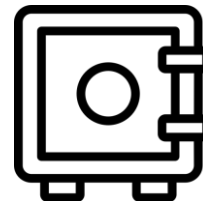
Objective: Procurement, processing, storage on ZBW repositories and long-term archiving of metadata and full texts

Subject: licenses on a national, an alliance or a consortium level as well as local licenses under the responsibility of the ZBW

Task:

- Clarification of necessary decisions
- Development of an IT infrastructure
- Development of workflows for processing the metadata and full texts of a license product

Start of the project: September 2021

End of the project: Archiving of the backlog and documentation of operational workflows

Afterwards: Archiving becomes a permanent task

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

…in the context of the mission of the ZBW

- World's largest research infrastructure for economic literature

- National library with a nationwide mission

- Part of the Leibniz Association (union of German non-university research institutes from various disciplines)

- Holdings: 4 million media units and access to 3.5 million electronic documents

- OA-Repository *EconStor*: > 250.000 publications in open access

- Repository *Digital Archive*: > 16,000 documents in open, restricted and closed access (closed access = SAVE)

- *EconBiz (*discovery system): > 12.4 million references from 9 sources

- Currently around 270 employees at two locations: Hamburg and Kiel

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Motivation I

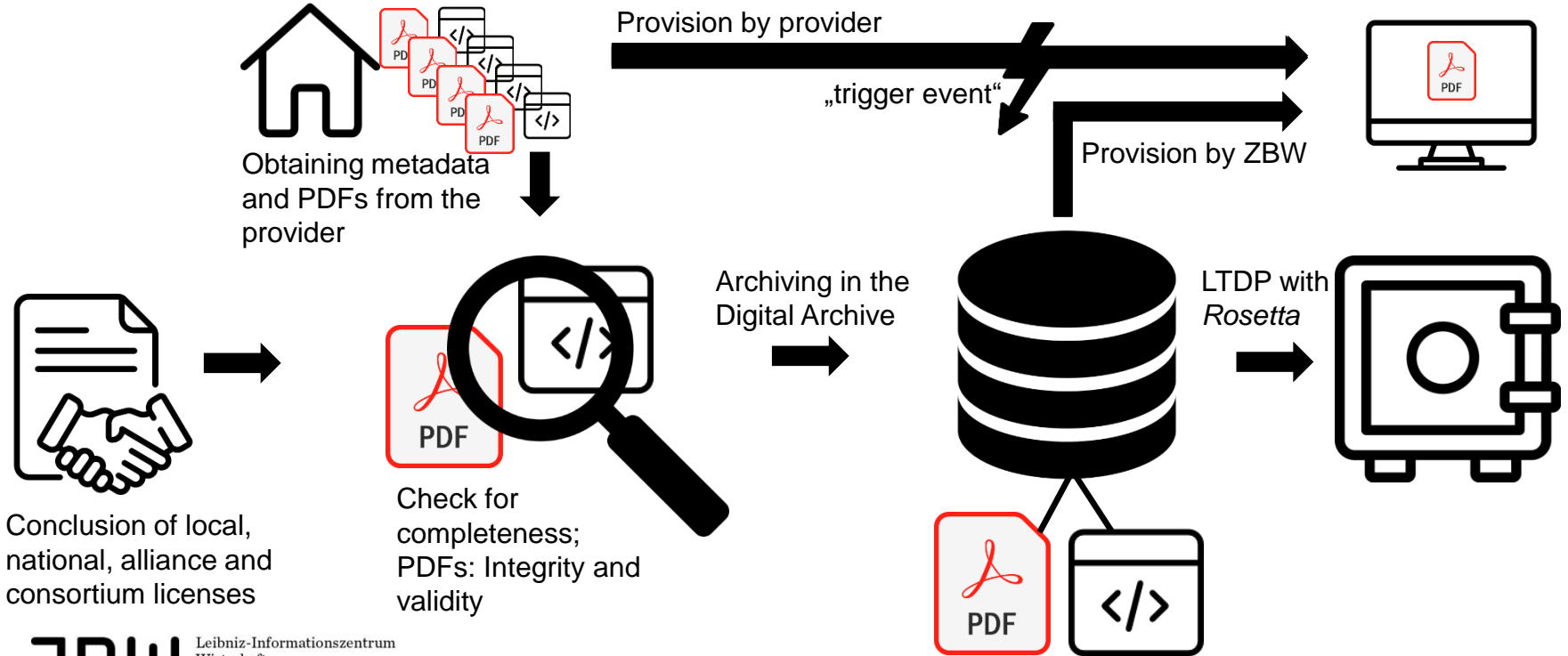The ZBW is committed to sustainability in the nationwide supply of literature

→ Archiving mandate as a central specialised library for supra-regional supply

    → Printed publications: permanent storage and preservation

    → also applies to digital resources

→ Contractual obligation to archive

    → to the DFG as funding organisation or

    → institutions participating in the consortium

Satzung der ZBW, § 2 Abs. 2

*„Die Stiftung sammelt und erschließt weltweit erscheinende wirtschaftswissenschaftliche Literatur. Sie bietet umfassende Serviceleistungen an, die eine **effiziente, effektive und nachhaltige** Nutzung wirtschaftswissenschaftlicher Fachinformationen ermöglichen.“*
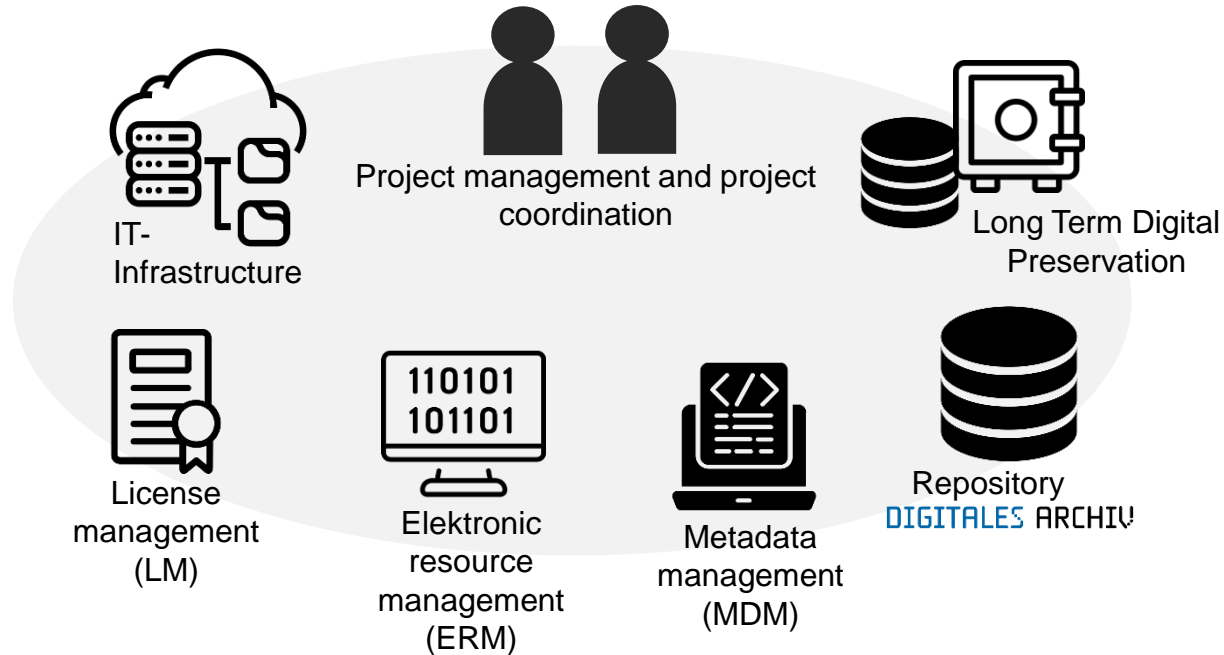
Motivation II

Obtaining metadata and PDFs from the provider

Provision by provider

„trigger event"

Provision by ZBW

Conclusion of local, national, alliance and consortium licenses

Check for completeness; PDFs: Integrity and validity

Archiving in the Digital Archive

LTDP with *Rosetta*

ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics

SAVE: Participants and areas of responsibility

IT-Infrastructure

Project management and project coordination

Long Term Digital Preservation

License management (LM)

Elektronic resource management (ERM)

Metadata management (MDM)

Repository
DIGITALES ARCHIV

Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

SAVE: IT Infrastructure

- Survey of the requirements for the necessary IT working environment in a joint discussion process with the IT infrastructure department
- Results:
    - FTP access for the delivery of data from the publishers
    - Server with 10 TB storage space with a virtual Linux machine
        - 2 storage pools:
            - Backup of the original data
            - For processing metadata and full texts
- React to increasing security requirements of publishers and by ZBW

Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

SAVE: Preliminary and ongoing work in cooperation with license management

Previous work and ongoing tasks:

- Complete listing and prioritisation of all licenses

- Clarification of the legal license framework (Archiving rights, data delivery), renegotiation if necessary

- License Package scope and annual growth → Storage space calculation

- Creating a project description for the acquisition of data from the providers, including our requirements

= Learning process, as the impact of contractually agreed rights is only now becoming visible with implementation

Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

SAVE: Acquisition of metadata and full texts

Start of the workflow: Completed licence negotiation

➢ LM establishes contact between provider and MDM

➢ MDM requests the metadata and full texts (PDFs) from the provider

➢ Retrieve the metadata and PDFs

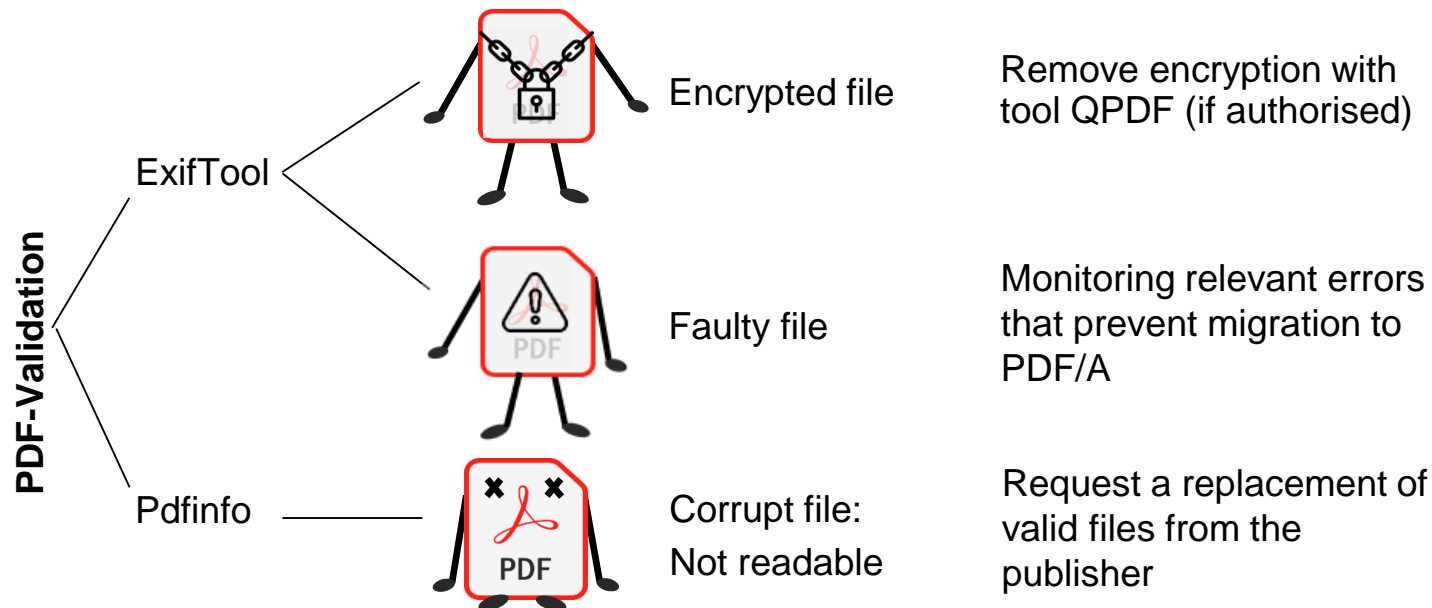➢ Storing on server of ZBW, automatic backups

Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

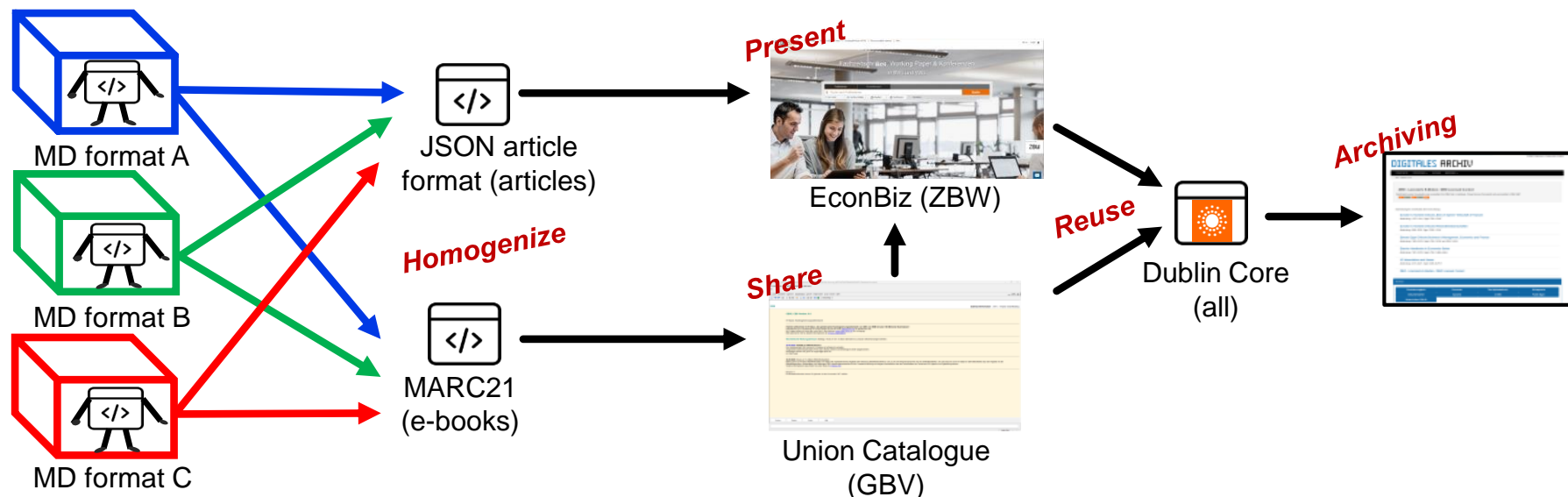# SAVE: Creating inventory lists and checking for completeness

- Pairwise match: PDF without metadata file, metadata file without PDF?

- Scope of the license based on the title list: Number of titles, volumes, issues, articles complete?

- Tools: Regular expressions in text editor, Linux command line tools, XML database management system using XPath/XQuery expressions

- Clarification with provider in case of discrepancies → request delivery of missing parts

→ iterative procedure until target and actual status are corresponding (as good as possible)

SAVE: Review of PDF files

PDF-Validation

ExifTool

Encrypted file — Remove encryption with tool QPDF (if authorised)

Faulty file — Monitoring relevant errors that prevent migration to PDF/A

Pdfinfo

Corrupt file: Not readable — Request a replacement of valid files from the publisher

SAVE: Flow of Metadata

Principles of workflow design: sustainability and efficiency
→ Reuse, homogenise, automate
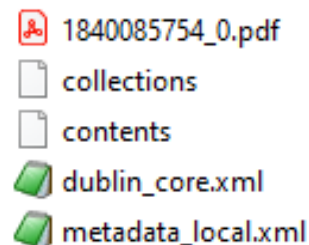
SAVE: Archiving in the repository DIGITALES ARCHIV

Software: **DSPACE**

- Defining the structures on the repository: Community, Collection, Item with Bitstreams (PDF)

- Setting up a closed access community (Dark Archive)

- Extension of the DSpace metadata schema (Dublin Core)

- Increasing storage space according calculation and main memory for better performance of the server

Upload process:

- Downloading metadata from the union catalogue (in MARC21)
  or from EconBiz (JSON article format)

- Conversion of the metadata into the target format Dublin Core

- Packing the import zips (PHP/Bash Script)

- Batch import via Linux server

1840085754_0.pdf
collections
contents
dublin_core.xml
metadata_local.xml

Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

SAVE: Archiving in the repository

DIGITALES ARCHIV

STARTSEITE   STATISTIKEN ▾   SUCHEN   BROWSEN ▾
ZBW - Digitales Archiv

ZBW - Lizenzierte E-Medien / ZBW Licensed Content
Restricted Access Documents only accessible from ZBW Kiel or Hamburg. Closed Access Documents only accessible to ZBW staff.
RSS 1.0   RSS 2.0   RSS

Sammlungen innerhalb der Einrichtung

Duncker & Humblot E-Books „Best of reprints" Wirtschaft & Finanzen
Abdeckung: 1875–1941, Sigel: ZDB-1-RWF

Duncker & Humblot E-Books Wirtschaftswissenschaften
Abdeckung 1996–2005, Sigel: ZDB-1-DHW

Edward Elgar E-Books Business & Management, Economics and Finance
Abdeckung: 1993–2022, Sigel: ZDB-1-EWE und ZDB-1-EEM

Elsevier Handbooks in Economics Series
Abdeckung: 1981–2022, Sigel: ZDB-1-HBE-article

US dissertations and theses
Abdeckung: 2015-2021, Sigel: ZDB-30-PDT

ZBW - Lizenzierte E-Medien / ZBW Licensed Content

DIGITALES ARCHIV

HOME   STATISTICS ▾   SEARCH   BROWSE BY ▾
ZBW - Digitales Archiv
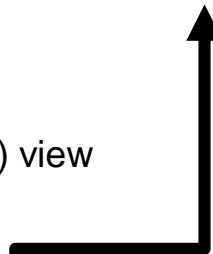
Authorization Required
You are not authorized to access the data.

Internal (ZBW) view

External view

SAVE: Digital Preservation in Rosetta

- in cooperation with the other two German National Libraries, the TIB - Leibniz Information Centre for Science and Technology University Library and the ZB MED - Information Centre for Life Sciences
- Rosetta software from Ex Libris company

- Long-term archiving = all measures that serve to permanently preserve digital objects for posterity (risk management, preservation planning, action, migration, etc.)

- Certified Digital Preservation at the ZBW: German nestor-Siegel (certification based on the DIN 31644 norm) in 2017 and Core Trust Seal in 2019

- End of workflow: License packages are transferred from the Digital Archive to Rosetta in an automated process → permanent preservation of metadata and full texts

- Rosetta is the digital preservation system for all ZBW document servers, such as EconStor OA repository, the digitisation centre and grey literature. Total number of records: over 500.000

ExLibris
Rosetta

ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics

Keepers: International provision of evidence of LTDP

Keepers:

- "The Keepers Registry acts as a global monitor on the archiving arrangements for continuing resources including e-serials." (https://keepers.issn.org/keepers-registry)
- Object: SAVE + OA repository EconStor + digitisation centre of the ZBW + grey literature
- Procedure to become a Keeper Archiving Agency
  - Filling out the Inclusion form
  - Creating a "characteristics" of the ZBW for integration into the portal
  - Conclusion of a MoU with the international ISSN Centre
  - Generate test data (standardised TSV file) and deliver to Keepers

Key-title Economic and Environmental Studies (Online)

⌃	Resource information		Archival Status				
Status	Publisher	Keeper	From	To	Updated	Extent of archive
Preserved	Uniwersytet Opolski	Internet Archive	2017	2020	18/12/2022	👁
Preserved	CEEOL	ZBW - Leibniz Information Centre for Economics	2012	2018	27/03/2023	👁

Keepers: Generation of the Keepers Report

Initial situation:

- (So far) no recording of LTDP in our union catalogue = no subsequent utilisation

Tasks:

- Generate the inventory information of the LTDP: Automated analysis of all data records with a handle of one of our repositories
- Generate the "Keepers Report": standardised TSV file

	A	B	C	D	E	F	G	H	I	J
1	eISSN	ISSN	Title of the resource	Publisher of the resource	Archive starting date	Archiv ending date	Status of archiving process	Holdings	Adress of the archived resource (URL or other)	Updating date
2	2304-1366		International journal of management, economics and social sciences : IJMES	Publisher unknown	2012	2022	Archived	1:1-11:1	https://hdl.handle.net/10419/62011	3/24/2023
3	1753-0296		The international journal of business science & applied management : IJBSA	Publisher unknown	2006	2022	Archived	1:1-17:1	https://hdl.handle.net/10419/190574	3/24/2023
4	2254-7088		European journal of government and economics : EJGE	Publisher unknown	2012	2019	Archived	1:1-2:1; 8:1-8:2	https://hdl.handle.net/10419/217703	3/24/2023
5	2148-8347		Journal of economics and political economy : JEPE	Publisher unknown	2016	2020	Archived	3:1-7:1	https://hdl.handle.net/11159/184	3/24/2023
6	2336-9205	1800-9581	Journal of central banking theory and practice	De Gruyter Open	2019	2020	Archived	8:1-9:1	https://hdl.handle.net/10419/217548	3/24/2023
7	2198-2627	2198-3402	Business research	Springer; VHB	2008	2020	Archived	1:1-13:3	https://hdl.handle.net/10419/103660	3/24/2023
8	2196-436X	2198-3526	Latin American economic review : LAER ; official journal of Centro de Invest	SpringerOpen	2014	2022	Archived	23:2014-31:2022	https://hdl.handle.net/10419/108948	3/24/2023
9	2197-893X	2568-3756	CO2-Barometer	KfW Bankengruppe	2013	2014	Archived	2013-2014	https://hdl.handle.net/10419/109737	3/24/2023
10	2568-3748	2568-3764	KfW/ZEW CO2 Barometer ...	KfW Bankengruppe; Centre for eu	2015	2016	Archived	2015-2016	https://hdl.handle.net/10419/109737	3/24/2023
11	2199-0026		CO2-Barometer	KfW-Bankengruppe	2013	2015	Archived	2013-2015	https://hdl.handle.net/10419/109737	3/24/2023
12	1869-0106	2195-6650	CO2-Barometer	KfW-Bankengruppe ; Zentrum fÃ¼	2009	2012	Archived	2009-2012	https://hdl.handle.net/10419/109737	3/24/2023
13	1867-6499	1867-5069	KfW/ZEW-GrÃ¼ndungspanel fÃ¼r Deutschland : BeschÃ¤ftigung, Finanzieru	Verb. der Vereine Creditreform	2008	2008	Archived	1:2008	https://hdl.handle.net/10419/109738	3/24/2023
14	1867-6499	1867-5069	KfW/ZEW-GrÃ¼ndungspanel	Verband der Vereine Creditreform	2009	2013	Archived	2:2009-6:2013	https://hdl.handle.net/10419/109738	3/24/2023
15	2365-8274	2365-8266	Mannheimer GrÃ¼ndungspanel	ZEW	2014	2015	Archived	2014-2015	https://hdl.handle.net/10419/109738	3/24/2023
16	1805-4846	1802-2197	European financial and accounting journal : EFAJ	Faculty of Finance and Accounting	2008	2020	Archived	3:1-15:2	https://hdl.handle.net/10419/109805	3/24/2023

ZBW
Leibniz-Informationszentrum
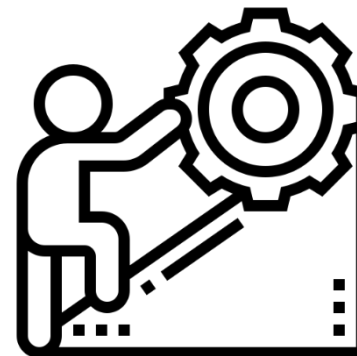Wirtschaft
Leibniz Information Centre
for Economics

Keepers: Generation of the Keepers Report

Challenges:

- Missing e-ISSNs: delivery not possible; cause: a lot of grey literature; in discussion: ISSN registration by ZBW via National ISSN Centre

- Cataloguing of p-ISSNs in the field for e-ISSNs: incorrect links in the Keepers portal

- Cataloguing rules of the journal database follow different principles than for ISSN assignment

- April 2023: Cyber attack; work had to be suspended ➔ Updates could not be created in the agreed timeframe, but hopefully at the end of this year

SAVE: Lessons learned

Adjustment of priorities regarding the processing of licensed products

- Initially focus on license types for which the ZBW has entered into obligations towards third parties: National, alliance, consortium licenses; local licenses downstream

- Reality check: Once communication has been established with a provider, all products licensed from the provider should be procured; communication about mutual expectations is very time-consuming and sometimes lengthy

Improved information flow between license management and metadata management

- Lack of clarity about the legal framework prevents quick decisions and generates a lot of effort. SAVE team is now involved in discussions between licence management and publishers at an early stage and has a certain say in the process

SAVE: Lessons learned

Better sooner than later

- Some contracts are old. Data was only requested long after the contract was signed. Publishers do not know what the ZBW wants to or should do with it ➔ Cover letter with motivation of the ZBW, profile and general conditions of the license as the start of communication with publishers.

- Now: involved in the process at an early stage, archiving follows directly after signing the contract

Automation potential in the workflow must be fully utilized

- Amount of data! Purchase of a product with >700,000 documents

- Update deliveries plus new license agreements accumulate the work running in parallel

Molehills grow to mountains: small sub-steps become major challenges

- e.g. Keepers: Generating the inventory data is much more time-consuming than expected during planning due to the existing framework conditions

- Tasks relating to metadata and archiving are difficult to plan due to various dependencies

ZBW
Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

Conclusion

Quite a lot of challenges, but worth doing it for sustainability and future generations of researchers and research

Tip for a virtual event

(in German)

- Participation is free of charge and registration is not required
- Presentations will be recorded and published in the TIB AV-Portal

What remains of formulas - a look at the practice of journal metadata
6 December 2023 : 14:40 - 15:00
Peter Eisner (TIB - Leibniz Information Centre for Science and Technology, Germany)

Titles and abstracts sometimes contain more than just text. Various approaches such as MathML or LaTeX are used to display formulae, for example. But what happens when this information encounters systems that can't do anything with it? Subject-specific requirements that go beyond simple text are already apparent in the title or abstract of scientific articles. For example, simple representations of molecules can occur (capital letters, subscript and superscript). But complex maths formulae are also not uncommon. In the old print world, such representations are unproblematic. In digital publications and retro-digitisation, however, they flow into the metadata. This means that fragments from XML or LaTeX sources encounter systems that sometimes only expect simple text information. Their definition context is lost in the process. Multiple conversions can result in errors and loss of information. In the TIB's digital long-term archiving team, the problem becomes apparent in the context of archiving open access journals, for which various metadata sources of sometimes very heterogeneous quality come into question. We take a look at examples from practice.
Link to the event

Conclusion

Quite a lot of challenges, but worth doing it for sustainability and future generations of researchers and research

Any questions?

Picture credits

- https://icon-icons.com/de/symbol/safe-box/12253123
- https://www.zbw.eu/fileadmin/images/press/standorte/wurr-01-hamburg-aussen.jpg
- https://www.zbw.eu/fileadmin/images/press/standorte/roth-02-kiel-aussen.jpg
- „Why?" Steve Rhodes (CC BY-NC-ND 2.0): https://www.flickr.com/photos/ari/3392824064/
- https://www.flaticon.com/de/kostenloses-icon/vertrag_26665
- https://icon-icons.com/de/symbol/safe-box/12253123
- https://freepikpsd.com/databasepng-transparent-images/419330/
- https://www.flaticon.com/premium-icon/metadata_3344335
- https://de.wikipedia.org/wiki/Datei:PDF_file_icon.svg
- https://www.flaticon.com/de/kostenloses-icon/vertrag_2666523
- https://www.freepik.com/icon/database_658099
- https://www.flaticon.com/de/kostenloses-icon/infrastruktur_3043474
- https://www.flaticon.com/de/kostenloses-icon/metadaten_6461809
  https://thenounproject.com/icon/binary-5389035/
- https://de.freepik.com/freie-ikonen/server_738376.html
- https://commons.wikimedia.org/wiki/File:Certificate_(89083)_-_The_Noun_Project.svg
- https://de.freepik.com/freie-ikonen/server_738376.html
- https://dspace.lyrasis.org/wp-content/uploads/2021/10/DSpace-logo.png
- https://avatars.githubusercontent.com/u/6927032?v=4
- https://keepers.issn.org/sites/all/themes/customkeepers/img/KeepersLogotype_BD.jpg
- https://thenounproject.com/icon/sustainability-2595668/
- https://www.vbib.net/
- https://thenounproject.com/icon/sustainability-2595668/

Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics